INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE

UMR 7501

Strasbourg

Thèse

présentée pour obtenir le grade de docteur de
l'Université de Strasbourg
Spécialité MATHÉMATIQUES APPLIQUÉES

Audrey Finkler

Modèle d'évolution avec dépendance au contexte et Corrections de statistiques d'adéquation en présence de zéros aléatoires

Soutenue le 16 juin 2010
devant la commission d'examen

Philippe Besse, rapporteur
Leonid Galtchouk, directeur de thèse
Catherine Matias, co-directrice de thèse
Christian Michel, examinateur
Jean-Marie Monnez, rapporteur
Photis Nobelis, invité

www-irma.u-strasbg.fr

cnrs dépasser les frontières

IRMA Institut de Recherche Mathématique Avancée

UNIVERSITÉ DE STRASBOURG

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE
Université de Strasbourg et C.N.R.S. (UMR 7501)
7 rue René Descartes
67084 STRASBOURG Cedex

Modèle d'évolution avec dépendance au contexte et Corrections de statistiques d'adéquation en présence de zéros aléatoires

par

Audrey Finkler

Mots-clés : modèles d'évolution des séquences d'ADN, processus de Markov, chaînes de Markov cachées, algorithme EM, tests d'hypothèses, tests d'adéquation, tables de contingence creuses, statistique du khi-deux, statistique de Kullback.

Classification Mathématique : 62F03 ; 62F05 ; 62M05 ; 62P10 ; 62P12 ; 92D15 ; 92D20 ; 92D30 ; 92D40.

Abstract

In this thesis we study the context-dependent evolution of DNA sequences.
In the first part we define a substitution model that not only distinguishes between transitions and transversions, but also allows for left-neighbor dependencies such as the CpG effect. We show that this model can be formulated as a hidden Markov model and we use the Baum-Welch algorithm to perform the parameter estimation. The model is then applied to estimate substitution rates observed in genetic sequences.
In the second part we develop corrections for classical goodness of fit test statistics with composite hypotheses for multinomial data in the presence of random zeros. Indeed, independence tests on the evolution of triplets of neighbor nucleotides involve contingency tables with numerous empty cells, and can be written as goodness of fit tests on sparse vectors. Thus, Pearson's and Kullback's statistics cannot be used. From these, we derive corrected statistics that share the same asymptotic behavior and apply these corrections to test independence on the evolution of nucleotide sequences. Finally, we propose applications to epidemiological and ecological data.

Keywords : models of DNA sequence evolution, Markov processes, hidden Markov models, EM algorithm, hypothesis testing, goodness of fit, sparse contingency tables, Pearson's statistic, Kullback's statistic.

Remerciements

Je voudrais tout d'abord remercier mon directeur de thèse, Leonid Galtchouk, et ma co-directrice de thèse, Catherine Matias, pour m'avoir permis d'entreprendre puis de mener à son terme un travail dont ce manuscrit est l'aboutissement. Merci à eux pour leur patience, leurs conseils et leur soutien.

Je suis très honorée que Philippe Besse et Jean-Marie Monnez aient accepté de rapporter ma thèse. Je suis également reconnaissante à Christian Michel pour sa participation à mon jury. Merci à eux pour leurs précieuses remarques et pour l'intérêt qu'ils ont porté à mon travail.

Je tiens également à exprimer toute ma gratitude à Photis Nobelis pour m'avoir fait partager ses intuitions et son expérience autour d'une problématique passionnante. Merci à lui pour sa gentillesse et sa disponibilité.

Que serait une thèse de mathématiques appliquées sans applications ? Merci à Isabelle Combroux, Mickaël Guedj et Carène Rizzon de m'avoir aidée dans la construction des jeux de données illustrant mon travail.

Je tiens à remercier chaleureusement Gilles Grasseau, Nicolas Klutchnikoff et Wolfgang Raffelsberger de m'avoir fait profiter de leur expertise en R, ainsi que Maurice Baudry, Pierre Navaro et Alain Sartout de m'avoir donné les moyens de mener à bien mes nombreuses et longues procédures informatiques. Merci également à Rémi et Adrien de m'avoir aidée à apprivoiser des animaux savants comme le gnou, le pingouin ou le saint-bernard.

Toute ma reconnaissance va aux personnels administratifs et techniques de l'université, partenaires aussi discrets qu'indispensables de notre réussite en tant qu'enseignants et doctorants, et garants de la qualité de vie du laboratoire. Merci donc à Claudine Bonnin, Yvonne Borell, Saïd Bouguerra, Sandrine Cerdan, Ferdaos Fassih, Fabienne Grauss, Daniel Grosson, Josiane Moreau, Grégory Thureau, et à beaucoup d'autres. Un grand merci également à Michèle Ilbert qui a fortement contribué à faciliter mes déplacements vers Evry, contribution appuyée par l'arrivée tant attendue du TGV Est ; je n'irai cependant pas jusqu'à remercier la SNCF pour ses retards et ses problèmes techniques...

Je tiens à remercier tous les membres, passés ou présents, du laboratoire Statistique et Génome pour leur accueil. Merci à Bernard Prum, Christophe Ambroise, Etienne Birmelé, Julien Chiquet, Pierre Latouche et Sophie Lèbre pour leur sympathie et leurs conseils.

Claudine Mitschi, Michèle Audin, Myriam et Frédéric Bertrand, Adeline Samson ainsi que mon tuteur pédagogique Michaël Gutnic m'ont toujours soutenue et entourée de leur bienveillance chaleureuse, et je leur en suis reconnaissante.

Merci à Anne-Laure, Guillaume et Ghislain pour leur amitié et leur soutien sans faille depuis une décennie. Merci également à mes camarades du célèbre bureau 113 et à tous les doctorants de l'IRMA. Ils contribuent à créer l'ambiance unique et stimulante régnant au laboratoire. Merci enfin à Nelly, Renaud, Denis, Christine, Yannick et Natha pour assurer l'ambiance en dehors du laboratoire !

Un immense merci à mes parents et ma grand-mère pour leur soutien inconditionnel et sans limite. Florian, merci pour tout, et pour plus encore...

Table des matières

Introduction **1**

Préambule : quelques notions de biologie moléculaire **3**

I Modèles d'évolution 9

1 Introduction aux modèles d'évolution **11**
1.1 Généralités 11
1.1.1 Modèles d'évolution 11
1.1.2 Propriétés générales des modèles 12
1.1.3 Fréquences stationnaires 14
1.1.4 Phylogénie et réversibilité 14
1.1.5 Indépendance 19
1.2 Modèles non uniformes et non homogènes avec indépendance 21
1.2.1 Évolution non uniforme 21
1.2.2 Évolution non homogène 22
1.3 Modèles avec dépendance au contexte 22
1.3.1 Chaînes de Markov cachées 23
1.3.2 Effet CpG 23
1.3.3 Généralisation aux modèles de n-uplets 24
1.3.4 Modèles avec dépendance au contexte pour séquences codantes 28

2 Modèle de quintuplets et calcul de vraisemblance **31**
2.1 Présentation du modèle de quintuplets 31
2.2 Trajectoires d'un processus markovien 33
2.3 Vraisemblance d'une trajectoire pour le modèle de quintuplets 38

3 Modèle de dépendance à gauche **43**
3.1 Description du modèle 43
3.1.1 Graphe des dépendances du modèle 43

3.1.2 Notre modèle vu comme une chaîne de Markov cachée 45
3.1.3 Paramètres du modèle 46
3.2 Estimation des paramètres 47
3.2.1 Algorithme Forward-Backward pour le calcul de l'espérance . 50
3.2.2 Étape de maximisation 52
3.3 Résultats 53
3.3.1 Paramètres 53
3.3.2 Identifiabilité du modèle 54
3.3.3 Simulations 56
3.3.4 Application à des données réelles 70
3.4 Perspectives 73

II Tests d'hypothèses pour tables de contingence creuses 75

4 Motivation biologique et notations 77
4.1 Motivation biologique 78
4.2 Notations pour les tables de contingence 79
4.3 Modèle multinomial 82
4.4 Tables et vecteurs creux 83

5 Statistiques d'adéquation 85
5.1 Statistiques de tests d'adéquation à des lois multinomiales 85
5.1.1 La famille des statistiques de divergence de puissance 86
5.1.2 Statistique du khi-deux de Pearson 87
5.1.3 Statistique de Kullback 89
5.1.4 Validité des tests et tables creuses 90
5.1.5 Test exact de Fisher 91
5.2 Correction de Ku pour la statistique de Kullback 92
5.2.1 Statistique de Kullback et tables creuses 92
5.2.2 Simulations 94
5.2.3 Cas d'un zéro unique 98
5.2.4 Cas de zéros multiples 100
5.2.5 Simulations 102
5.3 Corrections des statistiques de Pearson et Kullback 102
5.3.1 Contexte et nécessité de nouvelles corrections 103
5.3.2 Choix des corrections 103
5.3.3 Convergence en loi des statistiques corrigées 108
5.3.4 Simulations 110

6 Applications **115**
6.1 Exemple en dimension trois : évolution de triplets de sites nucléiques 116
6.2 Exemples en dimension deux 117
6.2.1 Approche multi-marqueurs pour la sclérodermie systémique . . 117
6.2.2 Trophie et végétaux des cours d'eau de la Petite Camargue Alsacienne 118
6.2.3 Perspectives 120

Bibliographie **121**

Table des figures

1 De l'ADN au chromosome par enroulements successifs 4
2 Une séquence de 10 pb et sa séquence complémentaire. 4

1.1 Arbre phylogénétique de protéines mitochondriales, extrait de [1]. . . 16
1.2 Arbre enraciné T_1 et arbre déraciné T_2 correspondant. 17
1.3 Substitutions à partir d'un dinucléotide CpG sur un brin et son complémentaire. 24

2.1 DAG du modèle de quintuplets. 32
2.2 Cône de dépendances pour le modèle de quintuplets. 33

3.1 DAG du modèle de dépendance au voisin de gauche. 44
3.2 DAG de la chaîne de Markov cachée. 45
3.3 Déroulement de l'algorithme Forward-Backward. 52
3.4 Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_2 et θ_6 pour $M = 3$ et pour des valeurs croissantes de n. 58
3.5 Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_4 et θ_7 pour $M = 3$ et pour des valeurs croissantes de n. 59
3.6 Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_2 et θ_6 pour $M = 4$ et pour des valeurs croissantes de n. 60
3.7 Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_4 et θ_7 pour $M = 4$ et pour des valeurs croissantes de n. 61
3.8 Boîtes à moustaches de 25 valeurs estimées des trois coordonnées de θ_2 et θ_6 pour $M = 5$ et pour des valeurs croissantes de n. 62
3.9 Boîtes à moustaches de 25 valeurs estimées des trois coordonnées de θ_4 et θ_7 pour $M = 5$ et pour des valeurs croissantes de n. 63
3.10 Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_1 et θ_2 pour $n = 1\,000$ et $M \in \{3, 4\}$, et de 25 valeurs estimées pour $M = 5$. 64
3.11 Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_3 et θ_4 pour $n = 1\,000$ et $M \in \{3, 4\}$, et de 25 valeurs estimées pour $M = 5$. 65

3.12 Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_5 et θ_6 pour $n = 1\,000$ et $M \in \{3, 4\}$, et de 25 valeurs estimées pour $M = 5$. 66
3.13 Boîtes à moustaches de 50 valeurs estimées des trois coordonnées du paramètre θ_7 pour $n = 1\,000$ et $M \in \{3, 4\}$, et de 25 valeurs estimées pour $M = 5$. 67
3.14 Arbre phylogénétique reliant l'Homme, le Chimpanzé et le Gorille. Les longueurs des branches sont données par les distances de Jukes-Cantor. 72

5.1 Quantiles d'ordre 0.95 de Q et G en fonction de c, sous les hypothèses nulles f_1 à f_6, pour 1 000 échantillons de $n = 400$ individus répartis dans $R = 100$ catégories. La ligne horizontale représente le seuil critique $\chi^2_{0.95,99}$. 97
5.2 Quantiles d'ordre 0.95 de Q, Q^{ab}, G, G^{ab}, G^{Ku} et $RC^{2/3}$ en fonction de c, sous les hypothèses nulles f_1 à f_3, pour 1 000 échantillons de 400 individus et $R = 100$ catégories. La ligne horizontale représente le seuil critique $\chi^2_{0.95,99}$. 112
5.3 Quantiles d'ordre 0.95 de Q, Q^{ab}, G, G^{ab}, G^{Ku} et $RC^{2/3}$ en fonction de c, sous les hypothèses nulles f_4 à f_6, pour 1 000 échantillons de 400 individus et $R = 100$ catégories. La ligne horizontale représente le seuil critique $\chi^2_{0.95,99}$. 113

Liste des tableaux

1 Code génétique. 5

3.1 Paramètres utilisés lors des simulations. 57
3.2 Moyennes $\overline{\mu}_{\mathrm{EM}}$, $\overline{\mu}_{\mathrm{KM}}$, $\overline{\mu}_{\mathrm{JC}}$, $\overline{\tau}_{\mathrm{EM}}$ et $\overline{\tau}_{\mathrm{KM}}$ calculées sur rep répétitions des estimations μ_{EM}, μ_{KM}, μ_{JC}, τ_{EM} et τ_{KM}, pour θ_2 et $M \in \{3, 4, 5\}$. 70

5.1 Probabilités multinomiales f_1 à f_6. 94
5.2 Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_1, selon c. . 95
5.3 Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_2, selon c. . 95
5.4 Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_3, selon c. . 95
5.5 Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_4, selon c. . 95
5.6 Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_5, selon c. . 95
5.7 Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_6, selon c. . 95
5.8 Comparaison des risques empiriques de première espèce pour les statistiques Q et G, pour 1 000 échantillons de $n = 400$ individus répartis dans $R = 100$ catégories, aux seuils 0.01, 0.05, 0.1, pour les lois multinomiales de probabilités f_1 à f_6 et le mode du nombre de zéros correspondant. 98
5.9 Comparaison des risques empiriques de première espèce pour les statistiques Q, Q^{ab}, G, G^{ab} et $RC^{2/3}$, pour 1 000 échantillons de $n = 400$ individus répartis dans $R = 100$ catégories, aux seuils 0.01, 0.05, 0.1, pour les lois multinomiales de probabilités f_1 à f_6 et le mode du nombre de zéros correspondant. 111
5.10 Probabilités multinomiales f_1^{bis} à f_6^{bis}. 114
5.11 Comparaison des puissances empiriques pour les statistiques Q, Q^{ab}, G, G^{ab} et $RC^{2/3}$, pour 1 000 échantillons de $n = 400$ individus répartis dans $R = 100$ catégories, au seuil 0.05 et pour le mode du nombre de zéros correspondant. Les vecteurs multinomiaux sont générés selon les probabilités f_1 à f_6 (Simulation). L'adéquation est testée pour les probabilités f_1^{bis} à f_6^{bis} ($\mathcal{H}_0^{\mathrm{bis}}$). 114

6.1 Table diplotypique pour l'étude de l'association entre trois marqueurs du gène $TNFAIP3$ et la sclérodermie systémique. 118
6.2 Table de contingence pour l'étude conjointe de la trophie et de la composition en espèces végétales. 119

Introduction

Depuis le développement des techniques de séquençage de l'ADN dans les années 1970, les biologistes disposent de bases de données très riches, dont le stockage et l'analyse nécessitent d'importants moyens informatiques ainsi que des modèles statistiques adaptés. Ces derniers permettent notamment une analyse « spatiale » qui consiste par exemple à localiser des gènes dans l'ensemble du matériel génétique, à analyser leur expression et à identifier leurs fonctions. Une analyse « temporelle » destinée à comparer des séquences soumises au processus d'évolution par mutation est également possible. Nous approfondissons ici cette seconde approche en restreignant les mutations possibles aux substitutions de nucléotides.

Les modèles d'évolution ont pour but de modéliser les divers types de mutations affectant l'ADN, et d'expliquer ainsi certaines différences fonctionnelles ou structurales entre les organismes. Ils permettent le calcul de distances évolutives et surtout la construction d'arbres phylogénétiques exhibant les relations de parenté entre espèces. Faciles à manipuler parce que décrits par un nombre limité de paramètres et aisément implémentables, les modèles évolutifs utilisés en pratique ne sont néanmoins pas satisfaisants pour les biologistes. Ces derniers remettent en effet en cause la plupart des hypothèses simplificatrices posées par les statisticiens. La levée progressive de ces hypothèses depuis les années 1990 nécessite le développement de méthodes de plus en plus complexes et numériquement coûteuses. Mon travail, composé de deux parties, se concentre sur la levée de l'hypothèse d'indépendance de l'évolution des sites d'une séquence.

La première partie est consacrée à l'étude de modèles d'évolution intégrant une dépendance des nucléotides à leur contexte. Après quelques rappels de biologie moléculaire figure dans le premier chapitre une présentation des principaux modèles de substitution développés jusqu'aujourd'hui. Dans un deuxième chapitre, nous décrivons un modèle récent avec dépendance au contexte, pour lequel nous détaillons le calcul de la vraisemblance d'une trajectoire. Ce modèle fait intervenir des regroupements de sites par quintuplets qui se chevauchent le long de la séquence. La complexité de la structure de dépendance rend la manipulation du modèle difficile. Ceci nous amène dans un troisième chapitre à considérer un nouveau modèle nucléotidique simple, qui prend en compte une dépendance temporelle markovienne unila-

térale. Plusieurs substitutions par étape évolutive sont admises et la dissymétrie de la dépendance rend le modèle non réversible. Nous n'observons pas tout le processus en temps continu, mais avons seulement accès aux extrémités de sa trajectoire. Nous développons ainsi le modèle pour des substitutions entre deux séquences connues et fixées, une séquence « ancêtre » et une séquence « descendant ». Par reformulation, ce modèle peut s'écrire sous la forme d'une chaîne de Markov cachée. L'estimation de ses paramètres est réalisée en maximisant itérativement la vraisemblance à l'aide d'une version de l'algorithme Expectation-Maximization spécifique aux chaînes de Markov cachées, appelée algorithme de Baum-Welch. La partie « Espérance » de l'algorithme comporte deux étapes : une étape de descente, aussi appelée « Forward », et une étape de remontée, aussi appelée « Backward ». Ces deux étapes se fondent sur les équations de filtrage, de prédiction et de lissage pour chaînes de Markov cachées. Nous avons, pour notre modèle, réécrit les équations Forward-Backward en exploitant sa structure particulière. L'étape M est effectuée numériquement. Le modèle est implémenté en langage R, et il est évalué à l'aide d'une série de simulations réalisées pour plusieurs jeux de paramètres. Ces derniers traduisent les principaux types de substitutions à l'œuvre dans les génomes, dont les transitions, les transversions, ainsi que les substitutions dues à l'effet CpG. Les simulations montrent que notre algorithme est performant dans l'estimation des paramètres. Nous appliquons enfin la procédure d'estimation à plusieurs séquences biologiques de caractéristiques différentes.

La problématique qui a induit la deuxième partie de mon travail découle de l'étude de la dépendance dans l'évolution des triplets de sites voisins d'une séquence d'ADN. Les tables issues des comptages de certains motifs sur des couples de séquences alignées présentent la particularité d'être creuses, c'est-à-dire qu'elles possèdent de nombreuses cases nulles. Nous considérons uniquement le cas où ces « zéros » sont aléatoires, c'est-à-dire qu'ils sont dus à un nombre d'observations trop faible. Les tables creuses mettent en défaut les tests d'indépendance classiques impliquant les statistiques du khi-deux de Pearson et du minimum d'information discriminante de Kullback. Nous traitons plus généralement le cas du test d'adéquation d'une hypothèse composée sur un échantillon de loi multinomiale dépendant de paramètres, et proposons des corrections pour les deux statistiques précédentes, adaptées au nombre de zéros. Nous montrons que les statistiques corrigées ont les mêmes propriétés asymptotiques que les statistiques dont elles sont issues. De plus, les tests dérivés sont aisément implémentables. Le risque empirique de première espèce est calculé par une méthode de Monte Carlo pour chacune des statistiques mises en jeu. Dans certains cas, nous évaluons par estimation la puissance de ces tests. Enfin, nous utilisons les statistiques corrigées pour effectuer des tests sur des données génomiques et, dans un cadre plus général, sur des données épidémiologiques et écologiques.

Préambule : quelques notions de biologie moléculaire

Nous introduisons dans ce préambule les principales notions de biologie moléculaire nécessaires à l'étude des modèles d'évolution. Ces notions sont détaillées dans le précis de génomique de Gibson et Muse [36].

Nucléotides

L'information génétique a pour support une longue molécule à double hélice appelée **ADN** (Acide DésoxyriboNucléique), localisée dans les cellules de chaque organisme vivant. Le monde du vivant se divise en deux grands domaines : les procaryotes, organismes unicellulaires sans noyau, et les eucaryotes, organismes plus complexes dont les cellules possèdent un noyau bien délimité.

Enroulée de façon complexe sur elle-même et, chez les eucaryotes, autour de perles appelées nucléosomes, cette molécule d'ADN constitue les chromosomes. Elle est composée d'une succession de quatre types de nucléotides représentés par la nature de leur base azotée : Adénine, Cytosine, Guanine et Thymine, notées respectivement $\mathbf{A}, \mathbf{C}, \mathbf{G}$ et $\mathbf{T}$. Les objets que nous considérons ici sont des séquences nucléotidiques, c'est-à-dire des portions d'ADN. Ce sont des réalisations $(x_1, \ldots, x_n)$ de suites finies $(X_1, \ldots, X_n)$ de variables aléatoires X_i à valeurs dans l'alphabet des bases $\mathcal{A} = \{A, C, G, T\}$, de cardinal $|\mathcal{A}| = 4$. Un site désigne un emplacement, une position de la séquence. Les dinucléotides sont des paires de bases adjacentes dans une séquence. Nous notons ainsi $X_i p X_{i+1}$ pour le couple (X_i, X_{i+1}) à valeurs dans $\mathcal{A}^2$, où p représente une liaison phosphodiester.

Les bases se divisent en deux classes selon leur structure chimique. Nous distinguons ainsi les **purines** $\mathcal{R} = \{\mathbf{A}, \mathbf{G}\}$ qui possèdent deux cycles azote-carbone, et les **pyrimidines** $\mathcal{Y} = \{\mathbf{C}, \mathbf{T}\}$ qui n'en possèdent qu'un.

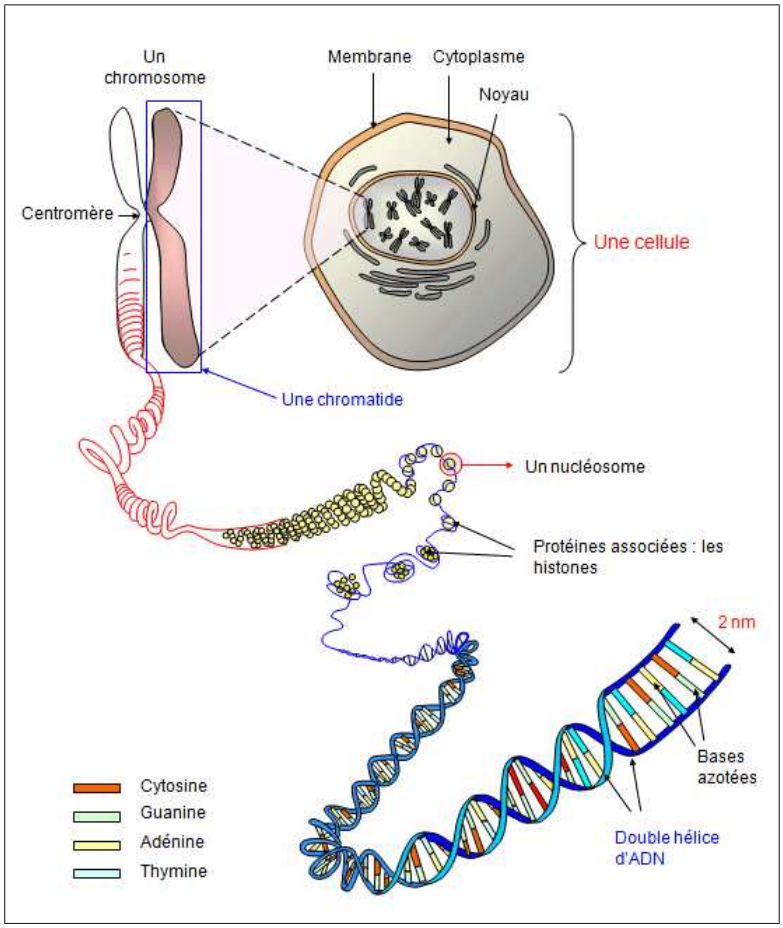


Figure 1 – De l'ADN au chromosome par enroulements successifs.
(image réalisée par Georges Dolisi)

$$\begin{array}{lcl} 5' & \overrightarrow{CAGTCGTATA} & 3' \\ & | \ | \ | \ | \ | \ | \ | \ | \ | \ | & \\ 3' & \underleftarrow{GTCAGCATAT} & 5' \end{array}$$

Figure 2 – Une séquence de 10 pb et sa séquence complémentaire.

Pour étudier la double hélice, il suffit en général de connaître un seul de ses deux brins parce que les bases sont chimiquement complémentaires au niveau de leurs liaisons hydrogène, liant toujours une Adénine à une Thymine, et une Cytosine à une Guanine. Les longueurs des séquences sont ainsi exprimées en paires de bases (**pb**). En raison des liaisons chimiques entre bases, les brins ont une extrémité **5′** et une extrémité **3′**, et un brin est lu de 5′ vers 3′.

Codons

Le dogme central de la biologie moléculaire chez les eucaryotes est résumé par le schéma suivant :

$$\text{ADN} \xrightarrow{\text{transcription}} \text{ARN} \xrightarrow{\text{épissage}} \text{ARNm} \xrightarrow{\text{traduction}} \text{protéine.}$$

La première étape, appelée **transcription**, conduit à la formation d'ARN (Acide RiboNucléique). Un **codon** est un groupe de trois nucléotides successifs d'une séquence d'ADN, d'ARN ou d'ARNm, formant une unité de sens. La troisième étape, appelée **traduction**, est contrôlée par le code génétique figurant dans la table 1, c'est-à-dire la correspondance redondante entre les $4^3 = 64$ codons possibles (quatre possibilités pour chacune des trois positions du codon) et les 20 acides aminés.

1 \ 2	**T**	**C**	**A**	**G**	3
	phénylalanine	sérine	tyrosine	cystéine	**T**
	phénylalanine	sérine	tyrosine	cystéine	**C**
T	leucine	sérine	STOP	STOP	**A**
	leucine	sérine	STOP	tryptophane	**G**
	leucine	proline	histidine	arginine	**T**
	leucine	proline	histidine	arginine	**C**
C	leucine	proline	glutamine	arginine	**A**
	leucine	proline	glutamine	arginine	**G**
	isoleucine	thréonine	asparagine	sérine	**T**
	isoleucine	thréonine	asparagine	sérine	**C**
A	isoleucine	thréonine	lysine	arginine	**A**
	méthionine	thréonine	lysine	arginine	**G**
	valine	alanine	acide aspartique	glycine	**T**
	valine	alanine	acide aspartique	glycine	**C**
G	valine	alanine	acide glutamique	glycine	**A**
	valine	alanine	acide glutamique	glycine	**G**

TABLE 1 – Code génétique.

Trois de ces codons engendrent un STOP, c'est-à-dire un signal d'arrêt de transcription, et sont appelés **codons-stop**. Deux codons sont dits **synonymes** lorsqu'ils codent pour le même acide aminé. C'est le cas notamment pour les codons CTC et CTA. La suite des codons permet de produire une suite d'acides aminés qui s'assemblent dans l'espace pour former une protéine. La plupart des protéines sont constituées de quelques centaines d'acides aminés. La fonction d'une protéine est déterminée par sa structure tridimensionnelle.

La deuxième étape, appelée **épissage**, est un processus consistant à éliminer les portions non codantes de l'ARN appelées **introns**. Les parties restantes, correspondant aux régions codantes appelées **exons**, sont soudées pour former l'ARN messager (ARNm). Une caractéristique importante de l'épissage est sa non-reproductibilité, c'est-à-dire que les coupes interviennent en fonction du contexte des processus biologiques en cours, et permettent ainsi de réguler l'expression des gènes. L'épissage est dit « alternatif », et constitue l'une des clés de l'épigénétique qui étudie les variations dans l'expression des gènes. Un gène humain contient en moyenne entre 10 et 15 exons et code ainsi au moins trois protéines différentes. Les portions d'ADN séparant les gènes constituent l'ADN intergénique.

Les données concernant la proportion de régions codantes, le nombre de gènes et plus précisément la composition en nucléotides varient d'une espèce à l'autre et d'une région nucléique à l'autre. De plus, elles sont régulièrement mises à jour. Ainsi, il y a encore 10 ans, le nombre de gènes chez l'Homme était estimé à 100 000, pour être réévalué depuis peu autour de 25 000. Une étude appelée « Human Genome Project », voir par exemple [78], précise que le génome humain se compose à 75% d'ADN intergénique, à 24% d'introns, et seulement à 1% d'exons.

Mutations

La **réplication** est le processus fiable mais non infaillible de duplication de l'ADN exploitant la complémentarité des brins, et il est indispensable au renouvellement cellulaire. Des mécanismes enzymatiques vérifient la qualité de la réplication, et assurent au besoin une correction des erreurs les plus graves. Celles qui échappent à ces corrections sont appelées mutations. Au fil du temps et à cause des mutations, les séquences évoluent et expliquent par là-même l'évolution génétique et phénotypique des espèces.

Une **substitution** est une mutation ponctuelle consistant à remplacer une base par une base différente en un court intervalle de temps. D'autres types de mutations existent comme :

- les **insertions** : introduction d'une nouvelle base dans la séquence,
- les **délétions** : suppression d'une base de la séquence,
- les **duplications** : copie puis insertion dans le génome d'une portion de la séquence,
- les **inversions** : renversement d'une portion de la séquence,
- les **transpositions** : déplacement d'une portion de la séquence.

Notons que ces trois derniers types de mutations sont des remaniements globaux, alors que les deux premiers peuvent être ponctuels, c'est-à-dire concerner un seul site. D'autre part, il existe trois cadres de lecture des codons selon que la transcription commence à la première, deuxième ou troisième lettre de la séquence, et

ces trois possibilités engendrent la plupart du temps des produits finaux de traduction différents. Les insertions et les délétions peuvent décaler le cadre de lecture et ainsi engendrer un STOP, ou rendre la protéine traduite incomplète voire non fonctionnelle.

Les substitutions pour lesquelles les deux nucléotides impliqués sont dans la même classe $\mathcal{R}$ ou $\mathcal{Y}$ sont appelées **transitions** et notées Ts, à l'inverse des **transversions**, notées Tv, qui induisent un changement de classe. Il y a deux fois plus de types possibles de transversions que de transitions, mais les transitions sont plus fréquentes que les transversions. En effet, le rapport entre le nombre de transitions ts et le nombre de transversions tv peut atteindre la valeur 10 ; il est en général supérieur à 1, comme montré dans [87].

Alignement de séquences

Les algorithmes d'alignement de séquences servent à positionner au mieux les sites de plusieurs séquences nucléiques ou protéiques les uns en-dessous des autres, pour que le plus grand nombre possible de positions de sites identiques coïncident. Les alignements permettent de reconstituer l'histoire évolutive des séquences, et les mutations qui ont eu lieu à partir de leur ancêtre commun. Comme la plupart des alignements ne peuvent être parfaitement ajustés, les algorithmes introduisent des trous dans les séquences, appelés **lacunes**, et qui sont représentés par un tiret pour chaque site manquant. Les insertions et délétions peuvent conduire à l'introduction de lacunes dans les alignements. Les algorithmes peuvent être paramétrés par l'utilisateur pour autoriser plus ou moins de lacunes selon la pénalisation qui leur est accordée. Plus de détails sur l'alignement de séquences figure dans l'ouvrage [31]. Dans tout ce travail, nous considérons ainsi des séquences ayant évolué uniquement par des substitutions, et issues d'alignements sans lacunes.

Première partie

Modèles d'évolution

Chapitre 1

Introduction aux modèles d'évolution

Les modèles d'évolution moléculaire sont principalement utilisés pour modéliser les processus de mutation, pour estimer des distances évolutives entre séquences et pour reconstruire des arbres phylogénétiques. Ces modèles peuvent également être utilisés pour étudier l'évolution de motifs nucléotidiques ou protéiques, et ainsi tester des hypothèses sur la structure des gènes primitifs, comme exposé dans [11].

Pour la phylogénie, il existe également des méthodes non probabilistes fondées sur le principe de parcimonie. Ce principe consiste à expliquer les différences observées à travers le moins possible de mutations. Nous nous intéressons ici à l'inverse aux modèles probabilistes, appliqués le plus souvent en pratique, et parmi eux aux **modèles de substitution markoviens à temps continu**. Les mutations autres que les substitutions ne sont pas considérées. Les modèles qui intègrent insertions et délétions existent et sont appelés modèles avec « indel », comme par exemple dans [58, 74, 75], mais ils ne sont pas étudiés ici.

1.1 Généralités

1.1.1 Modèles d'évolution

Le processus impliqué dans les substitutions et agissant sur les n sites d'une séquence $(X_1, \ldots, X_n)$ est supposé :

- **markovien** : la connaissance du passé du processus n'apporte pas d'information supplémentaire à la prédiction du futur si le présent est connu,
- **irréductible** : le processus permet d'atteindre tous les états de l'alphabet,
- **homogène** : les taux de substitution ne dépendent que de l'intervalle de temps séparant deux changements et non du moment où ils interviennent,
- **uniforme** : les sites de la séquence considérée évoluent de façon identique. Ainsi, les taux de substitution ne dépendent pas du lieu de la mutation.

Nous nous intéressons tout particulièrement à une dernière hypothèse, souvent adoptée parce qu'elle simplifie beaucoup les calculs théoriques, selon laquelle les sites évoluent de façon **indépendante** les uns par rapport aux autres.

1.1.2 Propriétés générales des modèles

Considérons plus généralement un espace probabilisé $(\Omega, \mathcal{P}, \mathbb{P})$ et un alphabet $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$ de $|\mathcal{A}|$ lettres. Les processus markoviens, irréductibles, homogènes et uniformes X_i agissant sur les bases sont caractérisés par une loi initiale π_0, et par les probabilités $\mathbb{P}(X_i(t) = y | X_i(0) = x)$ de substitution de la base x par la base y pendant un intervalle de temps de durée t. L'aspect markovien du processus se traduit par la **propriété de Markov faible** :

Définition 1. *Le processus X_i vérifie la propriété de Markov faible lorsque pour tout $h > 0$, nous avons :*

$$\mathbb{P}(X_i(t+h) = y | X_i(s) = x(s), s \leqslant t) = \mathbb{P}(X_i(t+h) = y | X_i(t) = x(t)). \tag{1.1.1}$$

Les probabilités $\mathbb{P}(X_i(t) = y | X_i(0) = x)$ forment la **matrice de transition** P_t d'éléments $P_t(x, y)$. Les coefficients de P_t sont indépendants du site i par uniformité, et de l'instant de substitution par homogénéité. Le processus vérifie l'équation suivante, obtenue en décomposant l'intervalle de temps en deux sous-intervalles, puis en conditionnant par tous les états intermédiaires possibles et enfin en utilisant la propriété de Markov (1.1.1).

Propriété 1. *La matrice de transition P_t du processus markovien vérifie l'égalité :*

$$P_{t+s} = P_t P_s,$$

appelée **équation de Chapman-Kolmogorov**, *et s'écrivant plus explicitement :*

$$P_{t+s}(x, y) = \sum_{z \in \mathcal{A}} P_t(x, z) P_s(z, y), \quad \forall s, t > 0, \forall (x, y, z) \in \mathcal{A}^3. \tag{1.1.2}$$

Définition 2. *La fonction matricielle $t \mapsto P_t$ est* **continue** *en probabilité en 0 si :*

$$\lim_{t \to 0} P_t(x, y) = \mathbb{1}_{x=y}, \quad \forall (x, y) \in \mathcal{A}^2.$$

L'équation (1.1.2) implique notamment les propriétés suivantes dont les énoncés et les preuves figurent dans l'ouvrage de Doob [28] :

- si le processus est continu en 0, alors il est continu en tout $t > 0$,
- si le processus est continu, alors pour t proche de 0, $P_t(x, x)$ est strictement positif pour tout x dans $\mathcal{A}$.

Nous supposons que le processus est continu en 0 et que $t \mapsto P_t$ est dérivable en tout $t \geqslant 0$. Nous définissons alors la matrice Q de taille $|\mathcal{A}| \times |\mathcal{A}|$, telle que :

$$\begin{aligned} Q_{x,y} &= \lim_{t\to 0} \frac{P_t(x,y)}{t} \quad \text{pour} \quad y \neq x, \\ Q_{x,x} &= -\lim_{t\to 0} \frac{1-P_t(x,x)}{t}. \end{aligned}$$

En différenciant l'équation (1.1.2) par rapport à s et à t puis en prenant respectivement $s = 0$ et $t = 0$, nous obtenons les **équations backward et forward de Kolmogorov**, respectivement (1.1.3) et (1.1.4), telles que pour tous x et y dans $\mathcal{A}$:

$$P'_t(x,y) = Q_{x,x}P_t(x,y) + \sum_{z\neq x} Q_{x,z}P_t(z,y), \tag{1.1.3}$$

$$P'_s(x,y) = P_s(x,y)Q_{y,y} + \sum_{z\neq y} P_t(x,z)Q_{z,y}. \tag{1.1.4}$$

Avec la condition initiale $P_0 = Id_{|\mathcal{A}|}$ où $Id_{|\mathcal{A}|}$ est la matrice identité de taille $|\mathcal{A}|$, la fonction matricielle P_t vérifie deux systèmes d'équations différentielles :

$$\begin{cases} P'_t = QP_t, \\ P_0 = Id_{|\mathcal{A}|} \end{cases} \quad \text{et} \quad \begin{cases} P'_t = P_tQ, \\ P_0 = Id_{|\mathcal{A}|}. \end{cases}$$

Il est classique de vérifier que ces équations possèdent une unique et même solution $P_t = e^{Qt}$, où Q est appelée **matrice de taux** ou **générateur infinitésimal** du processus. Les éléments diagonaux de Q sont définis de telle sorte que les sommes des lignes de Q valent zéro, et sont désignés par le symbole $*$.

Le calcul de P_t peut être réalisé à partir de Q principalement par deux méthodes. Premièrement, en considérant les premiers termes du développement de l'exponentielle matricielle pour k fixé :

$$e^{Qt} \simeq Id_{|\mathcal{A}|} + tQ + \frac{t^2}{2}Q^2 + \cdots + \frac{t^k}{k!}Q^k.$$

Notons qu'à l'ordre 1 cette approximation donne pour t proche de 0 :

$$P_t(x,y) \simeq \begin{cases} tQ_{x,y} & \text{si } x \neq y, \\ 1 + tQ_{x,x} & \text{si } x = y. \end{cases}$$

La deuxième méthode consiste à diagonaliser la matrice Q. Ainsi, $Q = SDS^{-1}$, d'où $P_t = Se^{Dt}S^{-1}$, avec D et e^{Dt} matrices diagonales, et S matrice inversible. Ces deux méthodes deviennent numériquement coûteuses lorsque $|\mathcal{A}|$ augmente.

1.1.3 Fréquences stationnaires

Définition 3. *Une mesure stationnaire du processus est une distribution de probabilité $\pi = (\pi_{a_1}, \ldots, \pi_{a_{|\mathcal{A}|}})$ solution de l'équation $\pi = \pi P_t$, pour tout $t > 0$. Cette mesure décrit la composition de la séquence à l'équilibre du processus.*

Théorème 1. *Si une chaîne de Markov irréductible et apériodique a un nombre fini d'états, alors elle possède une unique mesure stationnaire π, et pour toute loi initiale π_0 et tout $t > 0$, $\pi_0 {P_t}^n \xrightarrow[n\to+\infty]{} \pi$.*

Nous considérons dans ce travail des processus **stationnaires**, c'est-à-dire ayant atteint leur mesure stationnaire. Si nous choisissons comme mesure initiale $\pi_0 = \pi$, alors le processus est dit lui-aussi stationnaire et :

$$\mathbb{P}(X_i(t) = x) = \pi_x, \quad \forall t \geqslant 0, \forall x \in \mathcal{A}.$$

Pour l'alphabet $\mathcal{A} = \{A, C, G, T\}$, les fréquences stationnaires des nucléotides varient selon la nature de la séquence considérée. Elles diffèrent ainsi entre séquences codantes et non codantes. Il existe également des zones riches en G et en C, appelées **isochores riches en G + C**, et des zones pauvres en $G + C$.

1.1.4 Phylogénie et réversibilité

1.1.4.1 Modèles réversibles

Les modèles possédant les propriétés précédentes d'irréductibilité, d'homogénéité, d'uniformité, d'indépendance et de stationnarité ont généralement peu de paramètres, ce qui les rend faciles à manipuler. Ils sont pour la plupart **réversibles**, c'est-à-dire qu'ils vérifient la propriété suivante :

$$\overline{P}_t(x, y) = P_t(x, y), \quad \forall t \geqslant 0, \forall (x, y) \in \mathcal{A}^2, \tag{1.1.5}$$

où $\overline{P}_t(x, y)$ désigne la probabilité de substitution de x par y pendant l'intervalle de temps t en renversant la flèche du temps :

$$\overline{P}_t(x, y) = \mathbb{P}(X_i(0) = y | X_i(t) = x).$$

Par la formule de Bayes, l'équation (1.1.5) est équivalente à :

$$\pi_x P_t(x, y) = \pi_y P_t(y, x) \quad \forall t \geqslant 0, \forall (x, y) \in \mathcal{A}^2.$$

Il est facile de montrer que cette condition est aussi équivalente à la suivante, appelée **équilibre détaillé** :

$$\pi_x Q_{x,y} = \pi_y Q_{y,x}, \quad \forall (x, y) \in \mathcal{A}^2. \tag{1.1.6}$$

Proposition 1. *Si le modèle est réversible et si Q est symétrique, alors π est équirépartie :*

$$\pi = \left(\frac{1}{|\mathcal{A}|}, \ldots, \frac{1}{|\mathcal{A}|} \right).$$

Plus de détails sur la réversibilité figurent dans l'ouvrage de Kelly [48].

1.1.4.2 Distance évolutive

Une manière d'évaluer la distance séparant deux espèces est de compter le nombre de différences entre leurs deux séquences, témoignant de substitutions. Ce nombre est appelé distance évolutive, et vaut $K = 2kt$, où k est le taux global de substitution par site par unité de temps, et t la durée depuis leur divergence à partir d'un ancêtre commun.

Le taux k est calculé à partir des coefficients de la matrice de taux de substitution, qu'il faut en pratique estimer par la méthode du maximum de vraisemblance à partir de couples de séquences observées. Des calculs explicites de k à partir de couples de séquences sont donnés dans [47, 50], et les résultats sont rappelés au paragraphe 3.3.3 du chapitre 3 dans le cas du modèle de Kimura.

Lorsque les deux séquences considérées sont très distantes, les valeurs de K sont souvent sous-estimées parce que les observations ne font pas apparaître des substitutions « cachées », comme :

- les substitutions multiples : nous observons $x \to z$, alors qu'en réalité il s'est produit $x \to y \to z$ pour (x, y, z) dans $\mathcal{A}^3$,
- les substitutions inverses : nous n'observons pas de changement, mais il s'est produit $x \to y \to x$ pour (x, y) dans $\mathcal{A}^2$.

Les distances évolutives sont utiles dans la construction d'arbres phylogénétiques.

1.1.4.3 Phylogénie

Les modèles d'évolution sont en étroite relation avec la phylogénie moléculaire, c'est-à-dire l'étude évolutive des espèces au travers des séquences. Les liens de parenté entre espèces sont représentés à l'aide d'un arbre phylogénétique dont les branches symbolisent des relations de descendance, comme illustré par la figure 1.1.

Un arbre binaire est noté $T = (\mathcal{T}, \theta)$, où $\mathcal{T}$ est la topologie de l'arbre et θ regroupe les longueurs de branches. Les méthodes de reconstruction phylogénétique utilisent des modèles d'évolution. Certaines supposent que la topologie est connue, et estiment uniquement les longueurs de branches, d'autres cherchent l'arbre T le plus probable parmi un certain nombre de configurations possibles. La méthode développée initialement en phylogénie est l'estimation par parcimonie : l'arbre choisi est celui qui explique l'évolution entre deux séquences avec le moins de substitutions possibles. Cette méthode est peu fiable, mais numériquement intéressante dans le cas

de séquences très éloignées ou qui mutent rapidement. Les méthodes de reconstruction récentes utilisent souvent le principe du maximum de vraisemblance. Pour un arbre et un modèle d'évolution donnés, nous pouvons calculer la vraisemblance de l'arbre. Les paramètres du modèle et l'arbre sont ensuite optimisés afin de maximiser la vraisemblance. Notons également que, tous les types de séquences n'évoluant pas à la même vitesse ni de la même façon, la phylogénie des espèces peut différer selon le choix des séquences utilisées pour la reconstruire.

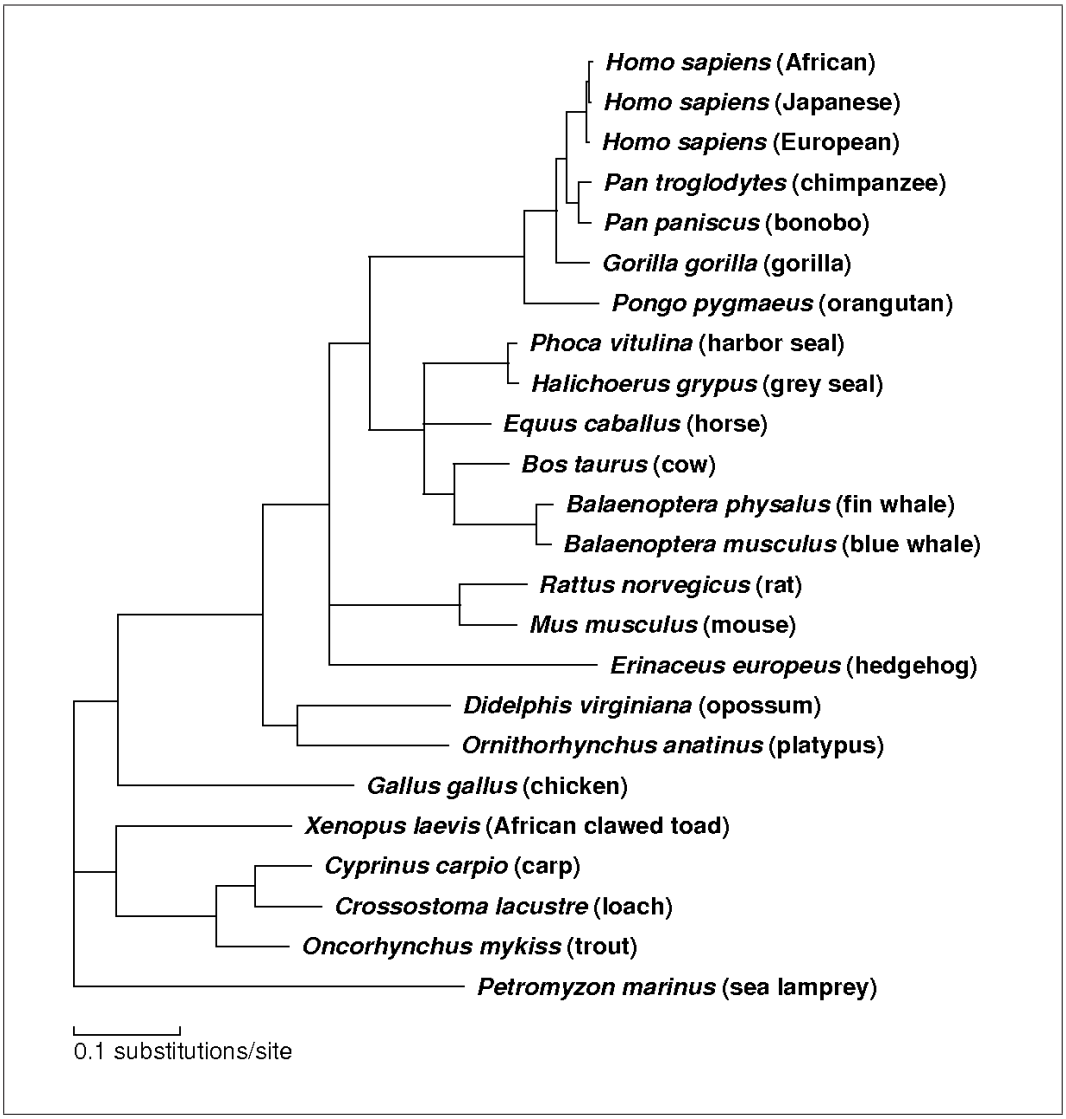


Figure 1.1 – Arbre phylogénétique de protéines mitochondriales, extrait de [1].

Felsenstein a proposé dans [30] une méthode d'inférence des arbres utilisant le principe du maximum de vraisemblance pour l'estimation des longueurs et un modèle d'évolution réversible en guise de support, implémentée ensuite dans le package PHYLIP du logiciel R (PHYLogeny Inference Package). Cette méthode a l'avantage de permettre des tests du type test du rapport de vraisemblance, comme dans [80], et sa justification biologique est plus évidente que celle de la parcimonie, mais elle est difficile à implémenter compte tenu du nombre énorme d'arbres à explorer. L'algorithme de Felsenstein balaie seulement un sous-ensemble de tous les arbres possibles en effectuant de petites perturbations à partir d'un arbre initial. Il propose également une méthode approchée consistant à ajouter les espèces à l'arbre au fur et à mesure de la procédure, en conservant à chaque étape la configuration donnant la vraisemblance la plus grande. L'arbre final dépend de l'ordre dans lequel les feuilles sont ajoutées, mais le résultat est une bonne approximation de l'arbre théorique. Des algorithmes plus puissants et plus rapides ont depuis été développés, notamment dans le package PHYML de R proposé dans [43], qui part d'un arbre complet et ajuste simultanément la topologie et la longueur des branches.

Si nous considérons deux séquences **homologues**, c'est-à-dire parentes, $\mathbf{x}$ et $\mathbf{y}$ de même longueur n séparées par la distance évolutive K, elles sont en réalité les feuilles d'un arbre T_1. Plus précisément, il existe une séquence $\mathbf{r}$ à la racine de l'arbre, ancêtre commun des séquences filles $\mathbf{x}$ et $\mathbf{y}$. Notons $t_{\mathbf{x}}$ et $t_{\mathbf{y}}$ les intervalles de temps séparant ces deux séquences de la racine $\mathbf{r}$, et posons $t = t_{\mathbf{x}} + t_{\mathbf{y}}$. L'arbre T_1 est représenté sur la figure 1.2 :

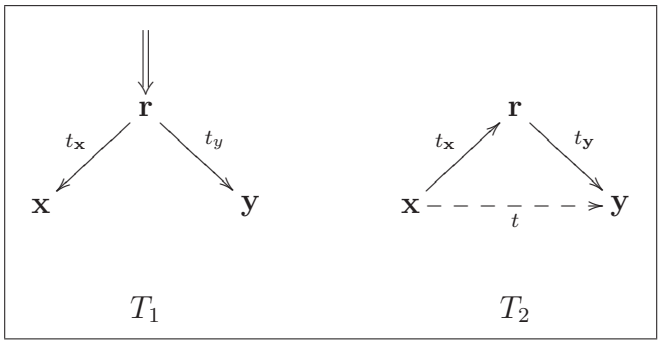


FIGURE 1.2 – Arbre enraciné T_1 et arbre déraciné T_2 correspondant.

Nous notons $P_t(\mathbf{x}, \mathbf{y})$ la probabilité de passage de $\mathbf{x}$ à $\mathbf{y}$ pendant l'intervalle de temps t quand $\mathbf{x}, \mathbf{y}$ sont des séquences de même longueur, et nous notons $\Pi_{\mathbf{x}}$ la loi stationnaire de la séquence $\mathbf{x}$. L'explicitation de ces quantités dépend du modèle évolutif utilisé. Lorsque l'évolution est supposée indépendante sur chaque site, les probabilités $P_t(\mathbf{x}, \mathbf{y})$ sont des produits de $P_t(x, y)$. Si en revanche un phénomène de dépendance est pris en compte dans l'évolution, l'écriture de $P_t(\mathbf{x}, \mathbf{y})$ est plus

complexe et fait intervenir une matrice des taux Q' adaptée aux transitions markoviennes entre séquences de longueur n : le nouvel alphabet $\mathcal{A}'$ est de taille 4^n, formé de toutes les séquences possibles de longueur n, écrites avec A, C, G et T. Quand n devient grand, la diagonalisation de Q' devient impossible à réaliser en raison des difficultés numériques mentionnées au paragraphe 1.1.2. Dans le cadre de la non-indépendance, il est donc important de trouver des méthodes n'utilisant pas ce type de calculs.

La vraisemblance du couple de séquences filles s'écrit :

$$L(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{r} \in \mathcal{A}^n} \Pi_{\mathbf{r}} P_{t_{\mathbf{x}}}(\mathbf{r}, \mathbf{x}) P_{t_{\mathbf{y}}}(\mathbf{r}, \mathbf{y}).$$

Ces calculs sont assez fastidieux, parce qu'il faut sommer sur toutes les racines possibles, mais ils se simplifient énormément à l'aide de l'équation (1.1.2) lorsque le modèle considéré est stationnaire et réversible, comme le souligne Felsenstein dans [30] :

$$\begin{aligned} L(\mathbf{x}, \mathbf{y}) &= \sum_{\mathbf{r} \in \mathcal{A}^n} \Pi_{\mathbf{r}} P_{t_{\mathbf{x}}}(\mathbf{r}, \mathbf{x}) P_{t_{\mathbf{y}}}(\mathbf{r}, \mathbf{y}), \\ &= \sum_{\mathbf{r} \in \mathcal{A}^n} \Pi_{\mathbf{x}} P_{t_{\mathbf{x}}}(\mathbf{x}, \mathbf{r}) P_{t_{\mathbf{y}}}(\mathbf{r}, \mathbf{y}), \\ &= \Pi_{\mathbf{x}} P_t(\mathbf{x}, \mathbf{y}). \end{aligned}$$

La vraisemblance $L(\mathbf{x}, \mathbf{y})$ est donc indépendante de la position de la racine qui est inconnue. Comme les taux sont homogènes et identiques sur les deux branches, l'horloge moléculaire est dite fixe, et la réversibilité permet d'inverser le sens des flèches en considérant $\mathbf{x}$ comme l'ancêtre de $\mathbf{y}$ ou réciproquement $\mathbf{y}$ comme l'ancêtre de $\mathbf{x}$. L'arbre T_1 est ainsi équivalent à l'arbre déraciné T_2, représenté en figure 1.2. Cette propriété a été nommée « principe de la poulie » dans [30]. Pour l'arbre enraciné T_1, imaginons que la racine $\mathbf{r}$ est une poulie qui peut glisser le long des branches. Tout arbre enraciné obtenu par un tel pivotement de T_1 autour de $\mathbf{r}$ est équivalent à l'arbre déraciné T_2. Signalons que la vraisemblance d'un arbre est généralement calculée à l'aide d'un algorithme qui part des feuilles de l'arbre complet, et regroupe les branches jusqu'à remonter à la racine, appelé algorithme d'élagage de Felsenstein.

Nous verrons que certains phénomènes biologiques témoignent de l'irréversibilité du processus d'évolution. Le statisticien a le choix d'imposer ou non la réversibilité à son modèle, en fonction des applications qu'il compte faire de celui-ci. En pratique, si le modèle est irréversible et la racine inconnue, il faut sommer sur toutes les racines possibles et si le modèle est réversible, il faut déterminer la position de la racine à l'aide d'une séquence extérieure au groupe phylogénétique considéré.

1.1.5 Indépendance

L'hypothèse selon laquelle les sites évoluent de façon indépendante les uns des autres est capitale. La vraisemblance conjointe de deux séquences $\mathbf{X} = (X_1, \ldots, X_n)$ et $\mathbf{Y} = (Y_1, \ldots, Y_n)$ s'écrit alors comme le produit des vraisemblances individuelles des sites :

$$\mathbb{P}_t(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = \Pi_{\mathbf{x}} \mathbb{P}_t(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} \pi_{x_i} P_t(x_i, y_i).$$

Il suffit alors de connaître π et la matrice des $P_t(x_i, y_i)$ pour pouvoir calculer cette vraisemblance.

Remarque 1. *Les matrices de taux de substitution seront données avec la convention suivante : les coordonnées correspondent aux bases dans l'ordre alphabétique* A, C, G, T.

Modèle de Jukes et Cantor Le plus simple de tous les modèles décrits précédemment est le modèle réversible de Jukes et Cantor, donné dans [47], qui suppose des taux tous égaux, et pour lequel le taux de substitution global k vaut 3α :

$$Q_{JC} = \begin{pmatrix} * & \alpha & \alpha & \alpha \\ \alpha & * & \alpha & \alpha \\ \alpha & \alpha & * & \alpha \\ \alpha & \alpha & \alpha & * \end{pmatrix}.$$

Modèle de Kimura Pour prendre en compte la déviation du taux ts/tv par rapport à 1, Kimura attribue dans son modèle présenté dans [50] un paramètre spécifique aux transitions, α, et un autre spécifique aux transversions, β :

$$Q_{K80} = \begin{pmatrix} * & \beta & \alpha & \beta \\ \beta & * & \beta & \alpha \\ \alpha & \beta & * & \beta \\ \beta & \alpha & \beta & * \end{pmatrix}.$$

Le taux global k est $\alpha + 2\beta$.

Ces deux modèles sont emboîtés : nous retrouvons le premier en fixant $\alpha = \beta$ dans le second. Grâce au test du rapport de vraisemblance utilisé dans [80], dont la statistique est distribuée asymptotiquement sous l'hypothèse nulle selon une loi du khi-deux, nous constatons que la vraisemblance des séquences augmente lorsque des paramètres ayant un sens biologique sont ajoutés. En revanche, même si le réalisme du modèle augmente, il est évident que la variance de K augmente également avec

le nombre de paramètres. Il est donc nécessaire d'employer une méthode de sélection de modèles pour déterminer celui qu'il est préférable d'utiliser en fonction du type de séquences considérées.

Les deux modèles présentés ci-dessus sont supposés stationnaires et vérifient $\pi_x = 1/4$ pour tout x dans $\mathcal{A}$. Les modèles suivants incluent l'hétérogénéité de la distribution des bases.

Modèle HKY Le modèle HKY de Hasegawa, Kishino et Yano présenté dans [44] prend en compte la distinction entre transitions et transversions. Les taux de substitution sont proportionnels à la fréquence stationnaire du nucléotide d'arrivée :

$$Q_{HKY} = \begin{pmatrix} * & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & * & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & * & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & * \end{pmatrix}.$$

Modèle GTR Le modèle « Rev » de Tavaré introduit dans [73] fait également intervenir les fréquences des bases :

$$Q_{Rev} = \begin{pmatrix} * & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & * & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & * & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & * \end{pmatrix}.$$

Il est aussi appelé modèle « GTR » pour General Time Reversible. C'est le modèle réversible le plus général possible, parce qu'il autorise des taux différents pour les six types de substitutions vérifiant l'équilibre détaillé (1.1.6).

Modèle Unrestricted Enfin, le modèle le plus général est le modèle « Unrestricted », sans hypothèse sur les taux :

$$Q_{Unres} = \begin{pmatrix} * & a & b & c \\ d & * & e & f \\ g & h & * & i \\ j & k & l & * \end{pmatrix}.$$

La comparaison de ces modèles emboîtés effectuée dans [83] montre que le modèle « Unrestricted » n'apporte que de petites améliorations, dont la portée ne compense pas l'inconvénient de l'augmentation de la variance des estimateurs des paramètres. Ainsi il est intéressant de considérer des modèles plus souples, et d'affaiblir les hypothèses simplificatrices posées initialement.

1.2 Modèles non uniformes et non homogènes avec indépendance

1.2.1 Évolution non uniforme

Parmi les hypothèses simplificatrices posées sur les modèles, la première à avoir été remise en cause est l'uniformité des taux de substitution le long de la séquence.

1.2.1.1 Non uniformité globale

Plusieurs catégories de sites alternent par plages le long des chromosomes : gènes et régions intergéniques, mais aussi exons et introns, respectivement codants et non codants, qui n'évoluent ni de la même façon, ni à la même vitesse. Le contenu en nucléotides peut également influer sur les taux de substitution. Il convient ainsi de distinguer les zones riches ou pauvres en $G + C$. Le contenu en nucléotides C et G est en effet lié à l'effet CpG que nous présentons au paragraphe 1.3.2.

Pour reconnaître des zones de distribution homogène dans l'ADN, une méthode très répandue depuis quelques années est l'analyse par chaînes de Markov cachées. Celle-ci permet d'identifier les plages successives, appelées **isochores**, à l'aide d'une méthode du type maximum de vraisemblance.

1.2.1.2 Non uniformité locale

De façon locale, c'est-à-dire à l'échelle de quelques nucléotides, il existe dans les régions codantes des variations dans les taux de substitution de positions voisines dans les codons. Lorsque nous observons la table du code génétique, il est clair que la troisième position est plus « libre » que les deux premières. Par exemple, les triplets CTT, CTC, CTA et CTG codent tous pour la leucine. Un changement à cette troisième position est souvent silencieux, c'est-à-dire que l'acide aminé produit par le codon et son codon muté est le même. Par conséquent il paraît raisonnable d'introduire des modèles de nucléotides distincts selon la position dans le codon, donc trois modèles munis chacun de paramètres à estimer.

En dehors des séquences codantes, les premières estimations du nombre de substitutions par site et par unité de temps supposent que les taux sont uniformes. Le processus évolutif étant markovien à temps continu, la variable aléatoire représentant ce comptage est distribuée suivant une loi de Poisson. Certains sites paraissent invariables, et les exclure améliore la qualité de l'ajustement à une loi de Poisson. Fitch et Markowitz suggèrent ainsi dans [34] que chaque site est durant toute son histoire soit dans un état « On » où il peut évoluer, soit dans un état « Off » où il est fixe. Ce modèle est appelé « Covarion » pour COncomitantly VARIable codONs.

De façon plus nuancée, Tuffley et Steel proposent dans [77] deux états possibles pour chaque site, donnant ainsi un modèle dit « Rates accross sites ». Dans un premier état, les sites évoluent de façon lente et dans un second, ils évoluent de façon plus rapide.

Pour créer des états intermédiaires à cette dichotomie, Yang a introduit dans [82, 84] des vitesses d'évolution prenant les k valeurs $r_1, \ldots, r_k$ d'une loi Gamma discrétisée, choisies pour chaque site par un tirage uniforme de probabilité $1/k$. Ces vitesses r_i sont des facteurs multiplicatifs de la matrice des taux, qui est commune. Yang choisit dans [84] comme référence le modèle HKY et montre, sur trois jeux de données de caractéristiques différentes, que l'amélioration la plus performante est obtenue pour k valant 4.

En réalité, les taux ne varient pas seulement dans l'espace, mais également dans le temps, le long des branches des arbres phylogénétiques.

1.2.2 Évolution non homogène

Goldman compare différents modèles de substitution dans [38, 79] grâce au test du rapport de vraisemblance, et y définit également des tests d'hypothèses permettant de comparer les vitesses d'évolution le long des branches d'un arbre. Il en conclut que les modèles autorisant les taux à varier en fonction de la branche sont en meilleure adéquation avec la réalité que les modèles à taux constant.

En dehors de leur modèle Rates accross sites, Tuffley et Steel étendent également dans [77] le modèle du Covarion de [34] à une évolution non homogène où les sites peuvent passer d'un état (On/Off) à l'autre au fil du temps. Ils alternent ainsi des périodes d'activité et de repos évolutifs, et cette alternance est régie par une chaîne de Markov à deux états.

Enfin, Galtier introduit dans [35] le modèle SSRV (Site Specific Rate Variation) qui généralise celui de Yang au cas où la vitesse d'évolution d'un site peut varier dans le temps en prenant k valeurs possibles.

De grands progrès dans l'adéquation entre modèles et réalité ont ainsi été réalisés. Ces modèles introduisent une plus grande flexibilité, mais supposent tous une évolution indépendante des sites les uns par rapport aux autres. Nous allons voir comment lever cette hypothèse également.

1.3 Modèles avec dépendance au contexte

L'abandon de l'indépendance des sites a permis un autre bond dans la qualité des approximations tout en engendrant une augmentation critique de la complexité des calculs et nécessitant des approches mathématiques et algorithmiques spécifiques. Les taux sont non seulement variables, mais aussi corrélés d'un site à l'autre.

Deux éléments viennent principalement réfuter l'indépendance de l'évolution entre les sites voisins d'une séquence. Le premier est une asymétrie dans la distribution des dinucléotides, en particulier pour le dinucléotide CpG, qui est la conséquence d'un biais mutationnel favorisant certaines substitutions. Le second est la structure des séquences codantes en codons. Nous traiterons ces deux points après avoir introduit les modèles à chaînes de Markov cachées.

1.3.1 Chaînes de Markov cachées

Yang définit dans [85] un modèle pour lequel la suite des taux de substitution de chaque site suit un processus markovien d'ordre 1 non observé (modèle en espace). Conditionnellement à ces taux, les sites sont indépendants et évoluent suivant le modèle HKY (modèle en temps). La combinaison de ces deux processus définit une chaîne de Markov cachée, notée **HMM** pour Hidden Markov Model, ou encore un modèle « espace-temps ». Les HMM sont très employées dans les modèles d'évolution parce qu'elles permettent de combiner un modèle spatial et un modèle temporel. Yang met en place un algorithme de type Expectation-Maximization (EM) pour chaines de Markov cachées pour l'estimation des paramètres. Nous détaillons cet algorithme au chapitre 3.

Jusqu'à présent, les modèles présentés supposaient que la phylogénie était connue et fixée pour les séquences données. Le modèle de Phylogenetic-HMM développé par Siepel et Haussler dans [72] pour décrire l'évolution de séquences codantes ou non codantes est la combinaison d'un processus markovien spatial le long du génome, qui attribue un arbre phylogénétique propre à chaque site, et d'un processus temporel le long des branches de cet arbre.

1.3.2 Effet CpG

Le phénomène chimique de méthylation-déamination des cytosines, mis en évidence par Bulmer dans [16], engendre chez les vertébrés un biais mutationnel qui se traduit par un déficit en dinucléotides CpG, aussi appelé **effet CpG**. En effet, les taux de substitution de CpG vers TpG, ou de CpG vers CpA sur le brin complémentaire, peuvent être dix à vingt fois supérieurs à ceux attendus.

Ces taux élevés engendrent une importante disparition des motifs CpG, dont la fréquence est donc moins élevée que celle attendue si les nucléotides évoluaient de façon indépendante. De façon générale, l'excès ou le déficit en un dinucléotide XpY est représenté par le rapport $XpY_{o/e}$, où figurent au numérateur le nombre observé n_{XpY} de motifs XpY, et au dénominateur le nombre théorique sous hypothèse d'indépendance $n_X.n_Y$ de motifs XpY :

$$XpY_{o/e} = \frac{n_{XpY}}{n_X.n_Y}.$$

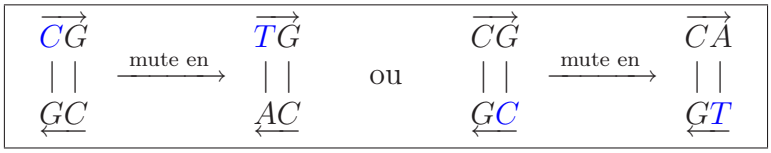
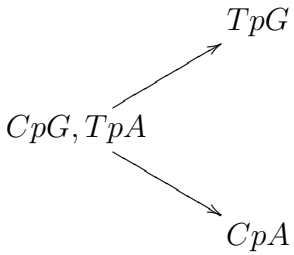


Figure 1.3 – Substitutions à partir d'un dinucléotide CpG sur un brin et son complémentaire.

Il existe cependant des régions du génome de quelques centaines ou quelques milliers de bases, appelées « îlots de CpG », qui sont moins touchées par ce phénomène du fait de contraintes de sélection et sont donc plus riches en CpG, comme exposé dans [5]. Parmi elles se trouvent certains exons et les régions promotrices de gènes constitutifs qui sont toujours exprimés. Une analyse de gènes et **pseudogènes** (gènes inactifs) humains menée dans [55] montre par exemple que la région promotrice de tous les gènes constitutifs et des gènes fortement exprimés contient un îlot de CpG.

Le même phénomène a lieu dans une moindre mesure à partir des dinucléotides TpA, voir [7], comme représenté sur le schéma suivant :



La cause du déficit en CpG et TpA dans certaines régions du génome est discutée dans [12, 29]. En effet, $CpG_{o/e}$ augmente dans les isochores riches en $G+C$ définis au paragraphe 1.1.3, alors que $TpA_{o/e}$ tend à diminuer dans ces régions.

Du point de vue évolutif, les modèles prenant en compte l'effet CpG sont irréversibles.

1.3.3 Généralisation aux modèles de n-uplets

Les n-uplets, aussi appelés **n-clusters**, sont des groupements de plusieurs sites adjacents qui se succèdent le long des séquences. Les plus utilisés sont les dinucléotides, groupements chevauchants ou non de deux sites voisins, et surtout les codons, groupements non chevauchants de trois nucléotides voisins codant pour des acides aminés. Deux codons voisins étant disjoints, ils constituent deux entités que nous pouvons facilement considérer comme indépendantes d'un point de vue statistique. Selon l'endroit où commence la lecture de la séquence, le découpage en triplets varie.

Les codons engendrent ainsi trois cadres de lecture possibles, comme expliqué dans le préambule.

1.3.3.1 Modèles de dinucléotides pour les régions non codantes

Les méthodes de dinucléotides étendent la réflexion sur les dinucléotides CpG et prennent en compte les occurrences de tous les couples de nucléotides se succédant en se chevauchant deux à deux le long de la séquence, comme dans [45]. L'objectif est de considérer des interactions simples entre sites voisins sans tomber dans l'écueil du cône de dépendances qui rend les calculs de vraisemblance inabordables, et dont une représentation figure à la section 2.1. Ce phénomène de cône de dépendances intervient lorsqu'un site dépend du passé de ses voisins, eux-mêmes dépendant du passé de leurs propres voisins, et ainsi de suite, étendant la portée des dépendances au cours de la remontée du temps. Pour cela, deux modèles non réversibles distincts ont été développés pour expliquer les substitutions entre deux séquences.

Modèles d'approximations de n-cluster Arndt, Burge et Hwa superposent dans [6] deux chaînes de Markov temporelles pour modéliser les substitutions : une chaîne de Markov simple qui agit comme dans les modèles avec indépendance sur les nucléotides, et une chaîne de Markov double qui agit sur les dinucléotides. Les fréquences des triplets sont estimées par les fréquences de dinucléotides $\hat{f}_{xyz} = f_{xy}f_{yz}/f_y$, provenant de l'hypothèse que la composition en nucléotides le long de la séquence est régie par une chaine de Markov d'ordre 1. Cette approximation est nommée approximation de 2-cluster. Lunter et Hein utilisent de façon similaire dans [57] un modèle de 3-cluster fondé sur une chaîne de Markov d'ordre 2 caractérisée par un générateur de taille 16×16. Pour l'estimation des paramètres du modèle, ils mettent en œuvre des méthodes « Markov Chain Monte Carlo » (MCMC) bayésiennes.

Dépendance à la classe $\mathcal{R}/\mathcal{Y}$ des voisins De leur côté, Bérard, Gouéré et Piau font intervenir dans [15] des substitutions dépendant de la classe purine ($\mathcal{R}$)/ pyrimidine ($\mathcal{Y}$) du nucléotide à gauche ou à droite du site considéré, en plus des taux de substitution simple habituels du modèle HKY, sous des contraintes d'égalité de certains taux de substitution. Cela leur permet de se départir du cône de dépendances sans faire d'approximations comme précédemment, et de calculer la loi stationnaire de façon exacte.

Voyons à présent comment modéliser les dépendances entre les codons, pour exploiter la structure en trinucléotides des séquences codantes.

1.3.3.2 Modèles de codons pour les régions codantes

Les modèles de codons sont bien appropriés aux séquences codantes, comme la séquence d'un gène après épissage par exemple. En effet, ils permettent d'intégrer au modèle la non-uniformité de la composition en nucléotides et une certaine forme de dépendance au contexte. Dans ces modèles, les éléments subissant les substitutions sont les codons, non plus les nucléotides, et une seule de leurs trois positions est modifiée par intervalle de temps. Les substitutions altérant plusieurs positions ont par conséquent un taux instantané nul.

Nous distinguons les mutations **synonymes** (ou silencieuses) des mutations **non synonymes**. Les premières ne changent pas la séquence d'acides aminés produite parce que le codon d'origine et le codon muté sont synonymes, ce qui est rendu possible par la redondance du code génétique, alors que la seconde provoque un changement dans la protéine et même parfois une interruption prématurée de sa synthèse par l'apparition d'un codon-stop. Le taux $\omega = dS/dN$ avec dS nombre de substitutions synonymes par site et par unité de temps et dN nombre de substitutions non-synonymes par site et par unité de temps, est un indicateur important de pression de sélection au niveau des protéines. Il peut être supérieur à 10, comme le montre [86].

Les modèles que nous présentons à présent permettent de prendre en compte des dépendances entre sites voisins au sein d'un codon, ainsi que des différences de taux d'évolution entre les positions dans le codon comme suggéré précédemment.

Modèle de codons de Muse et Gaut Le premier modèle de codons est celui de Muse et Gaut présenté dans [60]. Les paramètres μ et ν permettent de distinguer les taux des substitutions synonymes et non synonymes. Nous désignons respectivement par $\mathbb{1}_S$ et $\mathbb{1}_N$ les indicatrices de substitutions synonymes et non synonymes. Le générateur qui agit au niveau des codons $\mathbf{x} = (x_1, x_2, x_3)$ et $\mathbf{y} = (y_1, y_2, y_3)$ de $\mathcal{A}^3$ pour un changement intervenant en position i est donné par :

$$Q_{\mathbf{x},\mathbf{y}} = \pi_{y_i} \cdot \mu^{\mathbb{1}_S} \cdot \nu^{\mathbb{1}_N}.$$

Pour ce modèle, nous supposons que la fréquence stationnaire $\Pi_{\mathbf{y}}$ du codon $\mathbf{y} = (y_1, y_2, y_3)$ est égale au produit des fréquences stationnaires des trois nucléotides qui le composent : $\Pi_{\mathbf{y}} = \pi_{y_1}\pi_{y_2}\pi_{y_3}$. Les fréquences nucléotidiques sont ici préférées parce qu'elles sont moins nombreuses que les fréquences de triplets. Le modèle est alors dit « pondéré par les fréquences de nucléotides ». Conscients de la non homogénéité du processus biologique dans tout un arbre phylogénétique, les auteurs estiment les taux sur chacune des branches de l'arbre.

Modèle de codons de Goldman et Yang Dans le modèle un peu plus élaboré donné dans [39], les taux impliquant une transition sont multipliés par le facteur

$\kappa = ts/tv$, et par $\exp(-d_{aa_{\mathbf{x}},aa_{\mathbf{y}}}/V)$. La quantité $aa_{\mathbf{x}}$ représente l'acide aminé codé par le codon $\mathbf{x}$ et $d_{aa_{\mathbf{x}},aa_{\mathbf{y}}}$ est la distance de Grantham entre les acides aminés $aa_{\mathbf{x}}$ et $aa_{\mathbf{y}}$, extraite de [41], et fondée sur certaines propriétés physico-chimiques des vingt acides aminés. Plus la distance $d_{aa_{\mathbf{x}},aa_{\mathbf{y}}}$ est grande et plus faible est la probabilité d'observer $\mathbf{x} \to \mathbf{y}$. Le paramètre V représente la variabilité du gène, et il est inversement proportionnel à $\omega = dS/dN$. Les taux $Q_{\mathbf{x},\mathbf{y}}$ s'écrivent :

$$Q_{\mathbf{x},\mathbf{y}} = Z \cdot \Pi_{\mathbf{y}} \cdot \exp(-d_{aa_{\mathbf{x}},aa_{\mathbf{y}}}/V) \cdot \kappa^{\mathbb{1}_{Ts}},$$

et ils sont cette fois pondérés par les fréquences de codons $\Pi_{\mathbf{y}}$. Enfin, le coefficient de normalisation Z est choisi de telle sorte que le taux moyen de substitution soit égal à 1, c'est-à-dire :

$$\sum_{\substack{\mathbf{x},\mathbf{y}\in\mathcal{A}^3 \\ \mathbf{x}\neq\mathbf{y}}} \Pi_{\mathbf{x}} Q_{\mathbf{x},\mathbf{y}} = -\sum_{\mathbf{x}\in\mathcal{A}^3} \Pi_{\mathbf{x}} Q_{\mathbf{x},\mathbf{x}} = 1.$$

Modèle de codons de dépression en CpG Le modèle de dépression en CpG exposé dans [66] fait intervenir, en raison de la non-uniformité à l'intérieur des codons, la fréquence $\pi^i_{y_i}$ du nucléotide $y_i \in \mathcal{A}$ en position $i \in \{1,2,3\}$ dans le codon généré par le changement :

$$Q_{\mathbf{x},\mathbf{y}} = \pi^i_{y_i} \cdot \alpha^{\mathbb{1}_{Ts}} \cdot \beta^{\mathbb{1}_{Tv}} \cdot f^{\mathbb{1}_S} \cdot \lambda^{\mathbb{1}_{\text{ perte de } CpG} - \mathbb{1}_{\text{ gain en } CpG}}.$$

Comme le montre l'écriture de Q, ce modèle fait la distinction d'une part entre transitions et transversions, comme celui de Muse et Gaut, et d'autre part entre substitutions synonymes et non synonymes, comme celui de Goldman et Yang. Il introduit en plus un paramètre λ lié aux variations de contenu en CpG. Si la quantité en CpG reste constante au cours d'une substitution, le paramètre λ n'intervient pas. De même, si λ est fixé à 1, l'effet CpG n'est pas pris en compte.

Modèle de trinucléotides pseudo-chaotique Enfin, Bahi et Michel développent dans [8] un modèle d'évolution des gènes avec des mutations « pseudo-chaotiques » qui dépendent du temps, et qui agissent sur les triplets de nucléotides. Selon ce modèle, à chaque instant t une partie seulement des 64 trinucléotides peut subir une substitution. La matrice des taux varie ainsi en fonction de t. De plus, des taux différents sont associés aux trois positions de chaque triplet. Ce modèle permet notamment d'étudier l'évolution des codes circulaires identifiés dans les gènes des eucaryotes et des procaryotes.

1.3.3.3 Modèle de dépression en CpG pour les quintuplets

Même si les effets du contexte décroissent avec la distance au site considéré comme exposé dans [59], il paraît raisonnable d'étendre la portée aux quatre plus proches voisins en considérant des quintuplets au lieu de codons, comme dans [66]. Le modèle qui y est décrit est le prolongement du modèle de dépression en CpG, mais est plus complexe. Ici apparaissent en effet les premières véritables difficultés liées à la dépendance : les quintuplets successifs centrés sur des codons se chevauchent et ne permettent pas d'écrire la vraisemblance du modèle de manière simple. Il faudrait considérer une matrice Q de taille $4^n \times 4^n$ permettant de passer d'une séquence de longueur n à une autre. Les paramètres sont ainsi estimés en supposant les codons indépendants, et la vraisemblance exacte n'est pas calculée. Ce modèle a cependant l'avantage de prendre en compte les couples CpG se trouvant à cheval sur deux codons, ce qui était impossible avec les modèles de codons. De plus l'approximation avec des codons indépendants est en général de bonne qualité. Il ouvre également la voie à des méthodes capables de surmonter les difficultés engendrées par le chevauchement.

1.3.4 Modèles avec dépendance au contexte pour séquences codantes

Nous distinguons les modèles suivants de ceux présentés précédemment parce qu'ils utilisent des outils différents et sont de plus en plus exigeants en capacités de calcul. Ces derniers ne se résument plus à approcher des matrices de substitution ni à décomposer des formules de vraisemblance sur les sites.

1.3.4.1 Modèles avec chevauchement de quintuplets

Jensen et Perdersen ont prolongé dans [46] l'étude du modèle de dépression en CpG pour les quintuplets, toujours pour des séquences codantes. L'évolution des codons dépend cette fois de leur contexte dans la séquence. Les auteurs s'intéressent plus particulièrement à la loi stationnaire du processus et à ses relations avec la mesure de Gibbs. La vraisemblance ne s'écrit plus comme produit de vraisemblances individuelles sur les sites et il est impossible d'estimer les paramètres directement avec la méthode du maximum de vraisemblance comme auparavant. Souvent, la surface de la fonction de vraisemblance est en effet si complexe - c'est un espace de dimension correspondant au nombre de paramètres incorporés - qu'il est impossible de localiser le ou les pics de maximum de vraisemblance de manière explicite. C'est pour cette raison qu'ils mettent en œuvre des méthodes de simulation par MCMC et échantillonneur de Gibbs.

Jensen et Pedersen adaptent dans [65] ce modèle à des données qui sont des alignements de deux séquences contenant un gène qui code pour deux protéines différentes selon le cadre de lecture.

1.3.4.2 Utilisation de la pseudo-vraisemblance

Besag a introduit dans [13] une fonction de pseudo-vraisemblance inspirée de la théorie des champs markoviens, qui est un produit sur tous les sites de la vraisemblance du site conditionnellement à ses voisins. Christensen reprend ce concept dans [20], et développe l'algorithme EM correspondant pour estimer la pseudo-vraisemblance pour un modèle réversible semblable au modèle de dépression en CpG. Lorsqu'une substitution intervient entre les temps $t = 0$ et $t = 1$, il suppose que le changement a en fait lieu en $t = 1/2$, ceci pour réduire le nombre de chemins possibles. Dans [19], il étend son modèle au cas non réversible de manière à prendre en compte l'effet CpG de façon plus réaliste, comme nous l'avons souligné précédemment.

Nous approfondissons dans le chapitre 2 le modèle de quintuplets. Nous y calculons la vraisemblance des trajectoires pour l'évolution entre deux séquences connues.

Chapitre 2

Modèle de quintuplets et calcul de vraisemblance

Pour les modèles d'évolution supposant une évolution indépendante dans les séquences, les sites sont des variables aléatoires indépendantes et identiquement distribuées. L'évolution dans le modèle est alors dite **indépendante du contexte**. Nous avons exposé au chapitre 1 les principales raisons pour lesquelles cette hypothèse d'indépendance n'était pas justifiée d'un point de vue biologique. Nous avons ensuite vu que la levée de cette hypothèse demandait la mise en œuvre de nouvelles méthodes probabilistes et statistiques, souvent fondées sur des approximations et des procédures numériquement coûteuses. Dans ce chapitre, nous illustrons ce point par l'étude du modèle de dépendance de quintuplets présenté au paragraphe 1.3.4.1. Nous supposons comme auparavant que la phylogénie sous-jacente est fixée et connue, et nous explicitons la vraisemblance d'une trajectoire du processus pour l'évolution depuis un ancêtre jusqu'à son descendant. Nous montrons que l'expression de cette trajectoire est complexe et difficile à exploiter.

2.1 Présentation du modèle de quintuplets

Suite à l'article de Pedersen, Wiuf et Christiansen [66], Jensen et Pedersen introduisent dans [46], puis approfondissent dans [65], un modèle de substitution à temps continu avec dépendance au contexte, impliquant les codons. Le processus markovien évolue par des sauts successifs qui sont les substitutions. Nous supposons comme au chapitre 1 que les substitutions simultanées en deux sites distincts sont impossibles.

Considérons une séquence $\mathbf{X}(t)$ de n codons, soit $3n$ nucléotides, à l'instant t :

$$\mathbf{X}(t) = (\mathbf{X}_1(t), \ldots, \mathbf{X}_n(t)),$$

entourée de deux codons connus et fixes dans le temps $\mathbf{x}_0$ et $\mathbf{x}_{n+1}$. Cette séquence évolue, entre les instants $t = 0$ et $t = 1$, depuis son état initial observé $\mathbf{x}(0)$ jusqu'à son état final, également observé, $\mathbf{x}(1)$. Notons $\mathbf{X}_i = (\mathbf{X}_i(t))_{t\in[0,1]}$ l'histoire du site i, encore appelée **trajectoire** du processus $\mathbf{X}_i$. Décrivons l'action du modèle d'évolution de quintuplets sur les codons, puis plus précisément sur les nucléotides.

Une manière à la fois simple et rigoureuse de manipuler des dépendances multiples entre variables aléatoires est de les représenter par un graphe orienté appelé **DAG** pour Directed Acyclic Graph, présenté dans [56]. Le vocabulaire des DAG est le même que celui des arbres (parent, enfant, descendant, ancêtre, nœud, etc), et ces graphes possèdent la propriété suivante appelée propriété de Markov locale orientée, figurant à la page 51 de l'ouvrage [56].

Propriété 2. *Une variable est indépendante de ses non-descendants conditionnellement à ses parents.*

Dans le modèle de quintuplets, chaque codon $\mathbf{X}_i$ dépend de son passé immédiat ainsi que de celui de ses deux voisins, comme représenté par le DAG de la figure 2.1 pour un temps $t > 0$:

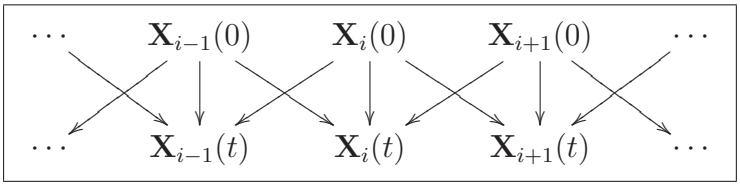


Figure 2.1 – DAG du modèle de quintuplets.

Désignons par X_i^j le j-ième nucléotide du codon i pour $j \in \{1, 2, 3\}$, et soit $\mathbf{X}_i^j = (X_i^{j-2}, X_i^{j-1}, X_i^j, X_i^{j+1}, X_i^{j+2})$ son **voisinage immédiat**. Ce voisinage donne son nom au modèle, parce qu'il se compose du site et de ses deux plus proches voisins à gauche et à droite, soit au total cinq nucléotides. Selon les cas :

$$\mathbf{X}_i^j = \begin{cases} (X_i^{-1}, X_i^0, X_i^1, X_i^2, X_i^3) = (X_{i-1}^2, X_{i-1}^3, X_i^1, X_i^2, X_i^3) & \text{si } j = 1, \\ (X_i^0, X_i^1, X_i^2, X_i^3, X_i^4) = (X_{i-1}^3, X_i^1, X_i^2, X_i^3, X_{i+1}^1) & \text{si } j = 2, \\ (X_i^1, X_i^2, X_i^3, X_i^4, X_i^5) = (X_i^1, X_i^2, X_i^3, X_{i+1}^1, X_{i+1}^2) & \text{si } j = 3. \end{cases} \tag{2.1.1}$$

La dépendance, non seulement à l'intérieur d'un codon, mais aussi aux deux nucléotides l'entourant, permet de modéliser les interactions entre dinucléotides, y compris à la frontière entre deux codons comme pour le modèle présenté dans [66]. Le fait de fixer les codons $\mathbf{x}_0$ et $\mathbf{x}_{n+1}$ aux extrémités spatiales permet de stopper la propagation des dépendances qui prend la forme d'un cône lorsque nous regardons le passé d'un codon. La figure 2.2 représente le cône de dépendances pour un codon ayant évolué en trois étapes dans une séquence de sept codons.

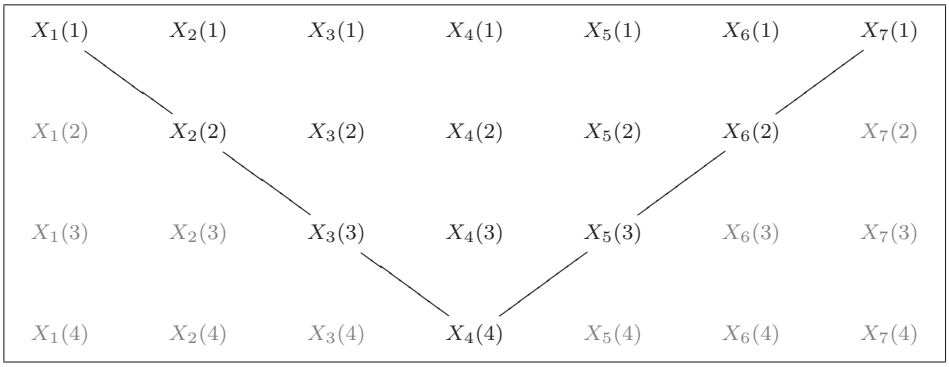


FIGURE 2.2 – Cône de dépendances pour le modèle de quintuplets.

Le chevauchement des quintuplets liés à deux codons voisins entraîne la dépendance de l'évolution des codons à leur voisinage. Il est donc impossible de décomposer le processus en sous-processus indépendants sur les quintuplets, et d'écrire la vraisemblance comme un produit de vraisemblances de quintuplets. Il faut en réalité étudier le processus dans son ensemble, et considérer qu'il agit sur la séquence entière comme expliqué dans le paragraphe 1.1.4.3 du chapitre 1. Nous considérons l'alphabet $\mathcal{A}$ des séquences de n codons non-stop. Ces codons sont au nombre de 61, car trois des 64 codons sont des codons-stop, comme le montre la table 1 du préambule, donnant $|\mathcal{A}| = 61^n$. Le générateur infinitésimal Q sur l'alphabet des séquences est donc de taille $61^n \times 61^n$.

Nous souhaitons connaître la vraisemblance d'une trajectoire conditionnellement à l'état initial $\mathbf{x}(0)$ de $\mathbf{X}$, notée :

$$L\left(\mathbf{x}(t), t \in [0, 1[\mid \mathbf{x}(0)\right),$$

c'est-à-dire la probabilité d'observer $\mathbf{x}(t)$ en connaissant son état initial. Pour cela, nous commençons par détailler la trajectoire d'un codon isolé soumis au processus de substitution markovien.

2.2 Trajectoires d'un processus markovien

Nous présentons ici la méthode du calcul de la vraisemblance des trajectoires d'un processus markovien Z à temps continu et à espace d'états fini $\mathcal{A}$, donnée par Albert dans [4]. Le processus est étudié pour $t \geqslant 0$ sur un espace probabilisé $(\Omega, \mathcal{T}, \mathbb{P})$. Sa trajectoire est notée $\{Z(t), t \geqslant 0\}$, et ses réalisations sont notées $\{Z(t, \omega), t \geqslant 0\}$ pour $\omega \in \Omega$.

Soit Q le générateur infinitésimal de Z, et $Q_x = -Q_{x,x}$ le taux de sortie de l'état x pour tout élément x de $\mathcal{A}$. Soit $\widetilde{Q}$ la matrice :

$$\widetilde{Q}_{x,y} = \begin{cases} Q_{x,y} & \text{si } x \neq y, \\ 0 & \text{si } x = y. \end{cases}$$

Nous utilisons dans la suite de ce chapitre les résultats du théorème suivant, exposés et démontrés au chapitre 6 de l'ouvrage [28].

Théorème 2. *Tout processus markovien Z à temps continu, à espace d'états fini, homogène et continu en 0 vérifie les propriétés suivantes :*

(i) Pour tous t_0 et α strictement positifs, la probabilité de rester pendant une durée α au point x en partant de x est :

$$\mathbb{P}\left(Z(t) = x,\ t_0 \leqslant t \leqslant t_0 + \alpha \mid Z(t_0) = x\right) = e^{-\alpha Q_x}.$$

(ii) Si $Z(t_0) = x$ et si le taux de sortie Q_x est strictement positif, alors presque sûrement il existe $t > t_0$ tel que $Z(t)$ soit une discontinuité dans la trajectoire de Z.

(iii) Soit $0 < \alpha \leqslant +\infty$ et t_1 l'instant de la première discontinuité ayant lieu dans l'intervalle $[t_0, t_0 + \alpha[$, si elle existe. Alors :

$$\mathbb{P}\left(Z(t_1) = y \mid Z(t_0) = x\right) = \frac{Q_{x,y}}{Q_x}.$$

(iν) Il existe un ensemble Ω' de Ω tel que $\mathbb{P}(\Omega') = 1$ et pour tout ω dans Ω',

$$Z(\cdot, \omega) : t \mapsto Z(t, \omega)$$

est une fonction en escalier avec un nombre fini de discontinuités sur tout intervalle de temps fini.

Les discontinuités sont également appelées **sauts**. La propriété (ii) du théorème 2 donne l'existence presque sûre de sauts pour des taux de sortie non nuls. Supposons que nous observons une trajectoire pendant une durée finie T. D'après la propriété $(i\nu)$, cette trajectoire $\{Z(t, \omega), t \in [0, T[\}$ est presque sûrement une fonction en escalier, caractérisée par le nombre de sauts dans $[0, T[$, par les intervalles de temps entre les sauts, ainsi que par les valeurs prises par le processus à l'issue de chaque saut. Soit K le nombre total de sauts. Pour $k \in \{1, \ldots, K\}$, appelons $\tau_k(\omega)$ l'instant d'occurrence du saut d'ordre k si $\omega \in \Omega'$, et posons $\tau_k(\omega) = +\infty$ pour $\omega \in \Omega \backslash \Omega'$. Posons également $\tau_0(\omega) = 0$. Pour $k \in \{0, \ldots, K\}$, les intervalles de temps séparant les sauts sont donnés par :

$$T_k(\omega) = \begin{cases} \tau_{k+1}(\omega) - \tau_k(\omega) & \text{si } \tau_{k+1}(\omega) < +\infty, \\ 0 & \text{sinon .} \end{cases}$$

Notons que $\sum_{k=0}^{K-1} T_k = \tau_K$. Soit également :

$$N(T,\omega) = \max_{k\in\mathbb{N},\tau_k(\omega)<T} \tau_k(\omega)$$

le nombre de sauts dans $[0,T[$, et :

$$Z_k(\omega) = Z(\tau_k(\omega),\omega)$$

l'état du processus juste après le saut d'ordre k pour k dans $\{0,\dots,K\}$. Avec ces notations, la trajectoire $\{Z(t,\omega), 0 \leqslant t < T\}$ est presque sûrement déterminée par le vecteur de dimension aléatoire à valeurs dans $\{\mathcal{A}\times[0,T[\}^{N(T,\omega)}\times\mathcal{A}$:

$$((Z_0(\omega),T_0(\omega)),\dots,(Z_{N(T,\omega)-1}(\omega),T_{N(T,\omega)-1}(\omega)),Z_{N(T,\omega)}(\omega)).$$

Écrivons à présent la fonction de répartition de cette trajectoire.

Théorème 3. *Pour le processus markovien Z, nous avons :*

$$\mathbb{P}(N(T)=0, Z_0=z_0) = \mathbb{P}(Z(0)=z_0)e^{-TQ_{z_0}}.$$

De plus, en fixant pour $K>0$ les notations $dt_0^{K-1} = dt_0\dots dt_{K-1}$ et

$$S_K = \{(t_0,t_1,\dots,t_{K-1}) : \sum_{k=0}^{K-1} t_k < T, 0\leqslant t_k \leqslant \alpha_k\},$$

nous avons également :

$$\begin{aligned}\mathbb{P}(N(T)=K, Z_0=z_0, T_0\leqslant\alpha_0,\dots,Z_{K-1}=z_{K-1},T_{K-1}\leqslant\alpha_{K-1},Z_K=z_K)\\ = \mathbb{P}(Z(0)=z_0)e^{-TQ_{z_K}}\int_{S_K}\prod_{k=0}^{K-1}\widetilde{Q}_{z_k,z_{k+1}}e^{-t_k(Q_{z_k}-Q_{z_K})}dt_0^{K-1}.\end{aligned}$$

Démonstration.
La première équation provient directement de la propriété (i) du théorème précédent. Démontrons la seconde équation. Prenons $K>0$ et notons $\widetilde{\mathcal{S}}_K$ l'événement $\{(T_0,T_1,\dots,T_{K-1})\in S_K\}$. Nous avons alors :

$$\begin{aligned}\mathbb{P}(N(T)=K, Z_0=z_0, T_0\leqslant\alpha_0,\dots,Z_{K-1}=z_{K-1},T_{K-1}\leqslant\alpha_{K-1},Z_K=z_K)\\ = \mathbb{P}\left(\left\{Z_0=z_0,\dots,Z_K=z_K,\sum_{k=0}^{K}T_k\geqslant T\right\}\cap\widetilde{\mathcal{S}}_K\right).\end{aligned}$$

Par le point (i) du théorème précédent et la propriété de Markov, nous avons :

$$\begin{aligned}\mathbb{P}(T_K > \alpha_K \mid Z_0 = z_0, \ldots, Z_K = z_K, T_0, \ldots, T_{K-1}) \\ &= \mathbb{P}\left(\left\{Z(t) = z_K, \sum_{k=0}^{K-1} T_k \leqslant t \leqslant \sum_{k=0}^{K-1} T_k + \alpha_K\right\} \;\middle|\; Z_K = z_K\right), \\ &= e^{-\alpha_K Q_{z_K}}.\end{aligned}$$

Ceci donne pour le complémentaire :

$$\begin{aligned}\mathbb{P}(T_K \leqslant \alpha_K \mid Z_0 = z_0, \ldots, Z_K = z_K, T_0, \ldots, T_{K-1}) \\ &= 1 - \mathbb{P}(T_K > \alpha_K \mid Z_0 = z_0, \ldots, Z_K = z_K, T_0, \ldots, T_{K-1}), \\ &= 1 - e^{-\alpha_K Q_{z_K}}.\end{aligned}$$

Comme la propriété markovienne implique :

$$\mathbb{P}(Z_K = z_K \mid Z_0 = z_0, \ldots, Z_{K-1} = z_{K-1}, T_0, \ldots, T_{K-1}) = \mathbb{P}(Z_K = z_K \mid Z_{K-1} = z_{K-1}, T_{K-1}),$$

cette quantité vaut $\widetilde{Q}_{z_{K-1},z_K}/Q_{z_{K-1}}$ par le point (iii) du théorème. Pour $k \geqslant 0$, considérons l'événement

$$\mathcal{R}_k = \{Z_0 = z_0, \ldots, Z_k = z_k, T_0 \leqslant \alpha_0, \ldots, T_k \leqslant \alpha_k\}.$$

Par conditionnements successifs, nous obtenons :

$$\begin{aligned}\mathbb{P}(\mathcal{R}_k) &= \mathbb{P}(T_k \leqslant \alpha_k, Z_k = z_k \mid \mathcal{R}_{k-1})\mathbb{P}(\mathcal{R}_{k-1}), \\ &= \mathbb{P}(T_k \leqslant \alpha_k \mid \mathcal{R}_{k-1}, Z_k = z_k)\mathbb{P}(Z_k = z_k \mid \mathcal{R}_{k-1})\mathbb{P}(\mathcal{R}_{k-1}), \\ &= (1 - e^{-\alpha_k Q_{z_k}})\frac{\widetilde{Q}_{z_{k-1},z_k}}{Q_{z_{k-1}}}\mathbb{P}(\mathcal{R}_{k-1}).\end{aligned}$$

Cette formule de récurrence nous donne :

$$\mathbb{P}(\mathcal{R}_K) = \mathbb{P}(Z(0) = z_0)\left[\prod_{k=0}^{K-1} \frac{\widetilde{Q}_{z_k,z_{k+1}}}{Q_{z_k}}\left(1 - e^{-\alpha_k Q_{z_k}}\right)\right]\left(1 - e^{-\alpha_K Q_{z_K}}\right),$$

$$= \mathbb{P}(Z(0) = z_0)\left[\prod_{k=0}^{K-1} \widetilde{Q}_{z_k,z_{k-1}} \int_{\substack{0 \leqslant t_k \leqslant \alpha_k \\ \forall k \in \{0,\ldots,K-1\}}} e^{-t_k Q_{z_k}} dt_0^{K-1}\right] \int_{0 \leqslant t_K \leqslant \alpha_K} Q_{z_K} e^{-t_K Q_{z_K}} dt_K.$$

Considérons la quantité suivante :

$$\mathbb{P}(Z_0 = z_0, \ldots, Z_K = z_K, T_0 \leqslant \alpha_0, \ldots, T_{K-1} \leqslant \alpha_{K-1}, T_K > \alpha_K) = \mathbb{P}(\mathcal{R}_{K-1}, Z_K = z_K, T_K > \alpha_K).$$

Par conditionnement, nous obtenons :

$$\mathbb{P}(\mathcal{R}_{K-1}, Z_K = z_K, T_K > \alpha_K) = \mathbb{P}(T_K > \alpha_K \mid \mathcal{R}_{K-1}, Z_K = z_K). \\ \mathbb{P}(Z_K = z_K \mid \mathcal{R}_{K-1})\mathbb{P}(\mathcal{R}_{K-1}),$$

qui vaut en utilisant le résultat précédent :

$$\mathbb{P}(Z(0) = z_0) \left[\prod_{k=0}^{K-1} \frac{\widetilde{Q}_{z_k, z_{k+1}}}{Q_{z_k}} \left(1 - e^{-\alpha_k Q_{z_k}}\right) \right] e^{-\alpha_K Q_{z_K}}.$$

Cette quantité est égale à :

$$\mathbb{P}(Z(0) = z_0) \left[\prod_{k=0}^{K-1} \widetilde{Q}_{z_k, z_{k+1}} \int_{\substack{0 \leqslant t_k \leqslant \alpha_k \\ \forall k \in \{0, \ldots, K-1\}}} e^{-t_k Q_{z_k}} \, dt_0^{K-1} \right] \int_{t_K > \alpha_K} Q_{z_K} e^{-t_K Q_{z_K}} \, dt_K.$$

Finalement, ceci correspond à la quantité recherchée :

$$\mathbb{P}\left(\left\{ Z_0 = z_0, \ldots, Z_K = z_K, \sum_{k=0}^{K} T_k \geqslant T \right\} \cap \widetilde{\mathcal{S}}_K \right) = \\ \mathbb{P}\left(\mathcal{R}_{K-1}, Z_K = z_K, T_K > T - \sum_{k=0}^{K-1} T_k \right),$$

qui vaut d'après les résultats précédents :

$$\begin{aligned} \mathbb{P}(Z(0) = z_0) \int_{S_K} \left\{ \int_{T - \sum_{k=0}^{K-1} t_k}^{\infty} Q_{z_K} e^{-t_K Q_{z_K}} \, dt_K \right\} \prod_{k=0}^{K-1} \widetilde{Q}_{z_k, z_{k+1}} e^{-t_k Q_{z_k}} \, dt_0^{K-1} \\ = \mathbb{P}(Z(0) = z_0) \int_{S_K} e^{-(T - \sum_{k=0}^{K-1} t_k) Q_{z_K}} \prod_{k=0}^{K-1} \widetilde{Q}_{z_k, z_{k+1}} e^{-t_k Q_{z_k}} \, dt_0^{K-1}, \\ = \mathbb{P}(Z(0) = z_0) e^{-T Q_{z_K}} \int_{S_K} \prod_{k=0}^{K-1} \widetilde{Q}_{z_k, z_{k+1}} e^{-t_k (Q_{z_k} - Q_{z_K})} dt_0^{K-1}. \end{aligned}$$

□

Les résultats de ce théorème nous permettent de calculer de façon explicite et sans approximation la vraisemblance de la trajectoire d'un codon pour le modèle de quintuplets. En effet, la quantité trouvée est une fonction de répartition qui, dérivée, nous donne l'expression de la densité du processus agissant sur chaque codon.

2.3 Vraisemblance d'une trajectoire pour le modèle de quintuplets

Appliquons les résultats théoriques précédents au calcul de la vraisemblance d'une trajectoire $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ de n codons $\mathbf{x}_i = (x_i^1, x_i^2, x_i^3)$ soumis au modèle de quintuplets. Pour cela, nous nous intéressons tout d'abord aux nucléotides. Nous faisons ensuite le lien entre les nucléotides et les codons, pour appliquer le théorème 3 aux trajectoires des codons, qui nous amènent enfin à la trajectoire d'une séquence de n codons.

Pour i dans $\{1, \ldots, n\}$ et j dans $\{1, 2, 3\}$, rappelons que le voisinage d'un nucléotide pour le modèle de quintuplets est donné par (2.1.1) et noté $\mathbf{x}_i^j$. Nous notons alors :

$$q\left(x_i^j, y_i^j \mid \mathbf{x}_i^j\right)$$

le taux instantané de substitution du nucléotide x_i^j en position j du codon i par le nucléotide y_i^j en fonction du contexte. Nous supposons que le processus de substitution est homogène et uniforme, donc ce taux dépend uniquement des nucléotides impliqués ; il ne dépend pas de l'instant de substitution ni de la position dans la séquence. Le taux de sortie de l'état x_i^j, c'est-à-dire le taux de substitution de x_i^j par n'importe quelle autre lettre de $\mathcal{A}$ s'écrit :

$$q\left(x_i^j \mid \mathbf{x}_i^j\right) = \sum_{\substack{y_i^j \in \mathcal{A} \\ y_i^j \neq x_i^j}} q\left(x_i^j, y_i^j \mid \mathbf{x}_i^j\right).$$

Considérons à présent deux séquences de codons $\mathbf{x} = (\mathbf{x}(1), \ldots, \mathbf{x}(n))$ et $\mathbf{y} = (\mathbf{y}(1), \ldots, \mathbf{y}(n))$ de même longueur n. La matrice de taux de transition Q du processus markovien est telle que $Q_{\mathbf{x},\mathbf{y}} = 0$ lorsque les séquences $\mathbf{x}$ et $\mathbf{y}$ diffèrent de plus d'un nucléotide. Ses éléments hors diagonale sont les taux de substitution nucléotidiques en fonction du contexte, c'est-à-dire :

$$Q_{\mathbf{x},\mathbf{y}} = \prod_{i=1}^{n} \prod_{j=1}^{3} q\left(x_i^j, y_i^j \mid \mathbf{x}_i^j\right).$$

En effet, les produits précédents ne contiennent en fait qu'un seul taux $q\left(x_i^j, y_i^j \mid \mathbf{x}_i^j\right)$, correspondant au codon subissant la substitution, et plus précisément aux deux nucléotides mis en jeu lors du saut, les autres taux étant tous égaux à 1. Le taux de sortie de la séquence $\mathbf{x}$ est la quantité suivante :

$$Q_{\mathbf{x}} = \sum_{i=1}^{n} \sum_{j=1}^{3} q\left(x_i^j \mid \mathbf{x}_i^j\right).$$

Notons, pour k dans $\{1,\dots,K\}$, τ_k les instants de substitution, $\mathbf{x}(\tau_k^-) = \mathbf{x}(\tau_{k-1})$ la séquence au moment précédant le saut d'ordre k, et $\mathbf{x}(\tau_k)$ la séquence juste après le saut d'ordre k. Les intervalles de temps entre les substitutions sont notés T_k. Nous montrons que le théorème 3 permet d'écrire la vraisemblance complète de $\mathbf{x}$ conditionnellement à son état initial. Soit K_{ij} le nombre de substitutions que subit la position j du site i au cours de son histoire. Alors :

$$\sum_{\substack{i\in\{1,\dots,n\}\\ j\in\{1,3\}}} K_{ij} = K.$$

Proposition 2. *La vraisemblance d'une trajectoire s'écrit :*

$$L\left(\mathbf{x}(t), t\in[0,1[\mid \mathbf{x}(0)\right) = \prod_{i=1}^{n}\prod_{j=1}^{3} e^{-\int_0^1 q\left(x_i^j(s) \mid \mathbf{x}_i^j(s)\right) ds} \cdot \prod_{k=1}^{K_{ij}} q\left(x_i^j(\tau_k^-), x_i^j(\tau_k) \mid \mathbf{x}_i^j(\tau_k^-)\right).$$

Démonstration.
La vraisemblance totale de $\mathbf{x}$ sur l'intervalle $[0,1[$ conditionnellement à la séquence initiale $\mathbf{x}(0)$ s'obtient en appliquant au processus de substitution avec dépendance aux quintuplets le résultat du théorème 3 pour $T=1$, et en le conditionnant par rapport à l'état initial. Il reste ainsi :

$$\begin{aligned} L &= L(\mathbf{x}(t), t\in[0,1[\mid \mathbf{x}(0)) \\ &= e^{-Q_{\mathbf{x}(\tau_K)}} \prod_{k=0}^{K-1} Q_{\mathbf{x}(\tau_k),\mathbf{x}(\tau_{k+1})} e^{-T_k(Q_{\mathbf{x}(\tau_k)} - Q_{\mathbf{x}(\tau_K)})}, \\ &= e^{-(1-\sum_{k=0}^{K-1} T_k)Q_{\mathbf{x}(\tau_K)}} \prod_{k=0}^{K-1} Q_{\mathbf{x}(\tau_k),\mathbf{x}(\tau_{k+1})} e^{-T_k Q_{\mathbf{x}(\tau_k)}}. \end{aligned}$$

Soient (i_0, j_0) les coordonnées du premier changement à partir de $\mathbf{x}(0)$. Remplaçons maintenant les taux par leur expression en fonction des quintuplets, ce qui donne :

$$L = \exp\left(-\sum_{i=1}^{n}\sum_{j=1}^{3} q\left(x_i^j(\tau_K) \mid \mathbf{x}_i^j(\tau_K)\right)\right) \exp\left(1 - \sum_{k=0}^{K-1} T_k\right) \cdot \prod_{i=1}^{n}\prod_{j=1}^{3}\prod_{k_{ij}=1}^{K_{ij}} q\left(x_i^j(\tau_{k_{ij}}^-), x_i^j(\tau_{k_{ij}}) \mid \mathbf{x}_i^j(\tau_{k_{ij}}^-)\right) \exp\left(-q\left(x_i^j(\tau_{k_{ij}}^-) \mid \mathbf{x}_i^j(\tau_{k_{ij}}^0)\right) T_{k_{ij}}\right),$$

puis

$$L = q\left(x_{i_0}^{j_0}(\tau_{k_{i_0j_0}}^-), x_{i_0}^{j_0}(\tau_{k_{i_0j_0}}) \mid \mathbf{x}_{i_0}^{j_0}(\tau_{k_{i_0j_0}}^-)\right) \exp\left(-q\left(x_{i_0}^{j_0}(\tau_{k_{i_0j_0}}^-) \mid \mathbf{x}_{i_0}^{j_0}(\tau_{k_{i_0j_0}}^-)\right) T_{k_{i_0j_0}}\right) \cdot$$
$$\prod_{i=1}^{n}\prod_{j=1}^{3}\left[\exp\left(-q\left(x_i^j(\tau_K) \mid \mathbf{x}_i^j(\tau_K)\right)\right) \exp\left(1 - \sum_{k=0}^{K-1} T_k\right) \cdot \right.$$
$$\left. \prod_{k_{ij}=1}^{K_{ij}} q\left(x_i^j(\tau_{k_{ij}}^-), x_i^j(\tau_{k_{ij}}) \mid \mathbf{x}_i^j(\tau_{k_{ij}}^-)\right) \exp\left(-q\left(x_i^j(\tau_{k_{ij}}^-) \mid \mathbf{x}_i^j(\tau_{k_{ij}}^-)\right) T_{k_{ij}}\right)\right].$$

Finalement,

$$L = L'^{j_0}_{i_0} \prod_{i=1}^{n}\prod_{j=1}^{3} L_i^j,$$

où

$$L'^{j_0}_{i_0} = q\left(x_{i_0}^{j_0}(\tau_{k_{i_0j_0}}^-), x_{i_0}^{j_0}(\tau_{k_{i_0j_0}}) \mid \mathbf{x}_{i_0}^{j_0}(\tau_{k_{i_0j_0}}^-)\right) \exp\left(-q(x_{i_0}^{j_0}(\tau_{k_{i_0j_0}}^-) \mid \mathbf{x}_{i_0}^{j_0}(\tau_{k_{i_0j_0}}^-))T_{k_{i_0j_0}}\right),$$

et

$$L_i^j = \exp\left(q\left(x_i^j(\tau_K) \mid \mathbf{x}_i^j(\tau_K)\right)\left(1 - \sum_{k=0}^{K-1} T_k\right)\right) \cdot$$
$$\prod_{k_{ij}=1}^{K_{ij}} \exp\left(-q\left(x_i^j(\tau_{k_{ij}}^-) \mid \mathbf{x}_i^j(\tau_{k_{ij}}^-)\right) T_{k_{ij}}\right) \prod_{k_{ij}=1}^{K_{ij}} q\left(x_i^j(\tau_{k_{ij}}^-), x_i^j(\tau_{k_{ij}}) \mid \mathbf{x}_i^j(\tau_{k_{ij}}^-)\right).$$

Nous retrouvons bien l'expression recherchée :

$$L = \prod_{i=1}^{n}\prod_{j=1}^{3} e^{-\int_0^1 q\left(x_i^j(s) \mid \mathbf{x}_i^j(s)\right) ds} \prod_{k_{ij}=1}^{K_{ij}} q\left(x_i^j(\tau_{k_{ij}}^-), x_i^j(\tau_{k_{ij}}) \mid \mathbf{x}_i^j(\tau_{k_{ij}}^-)\right), \quad (2.3.1)$$

parce que l'intégrale ci-dessus s'écrit de façon équivalente :

$$\sum_{k_{ij}=1}^{K_{ij}} \int_{\tau_{k_{ij}}}^{\tau_{k_{ij}+1}} q\left(x_i^j(\tau_{k_{ij}}) \mid \mathbf{x}_i^j(\tau_{k_{ij}})\right) ds + q(j \mid \mathbf{x}_i^j(\tau_K)\right)\left(1 - \sum_{k=0}^{K-1} T_k\right),$$

soit

$$\sum_{k_{ij}=1}^{K_{ij}} q\left(j \mid \mathbf{x}_i^j(\tau_{k_{ij}})\right)(\tau_{k_{ij}+1} - \tau_{k_{ij}}) + q\left(j \mid \mathbf{x}_i^j(\tau_K)\right)\left(1 - \sum_{k=0}^{K-1} T_k\right).$$

Ceci nous permet d'écrire la première partie de (2.3.1) sous la forme :

$$e^{-q\left(j \mid \mathbf{x}_i^j(\tau_K)\right)\left(1-\sum_{k=0}^{K-1} T_k\right)} \prod_{k_{ij}=1}^{K_{ij}} e^{-q\left(j \mid \mathbf{x}_i^j(\tau_{k_{ij}})\right)T_{k_{ij}}},$$

ce qui termine cette preuve. □

Nous obtenons ainsi une expression de la vraisemblance d'une trajectoire pour le modèle de quintuplets, dépendant des intervalles de temps séparant les sauts, ainsi que des valeurs prises par les sites au fil des substitutions. Lorsque nous considérons un problème d'estimation, nous nous intéressons à la somme des vraisemblances de toutes les trajectoires possibles du processus. La somme des expressions précédentes sur toutes les trajectoires possibles est très difficile à manipuler, même lorsque le temps est discrétisé. C'est pourquoi nous proposons un modèle avec dépendance au contexte plus simple. Ce modèle, présenté dans le chapitre 3, inclut les principales influences du voisinage sur les sites, et est plus aisé à manipuler du point de vue statistique.

Chapitre 3

Modèle de dépendance à gauche

Les principales difficultés rencontrées dans les modèles de n-uplets sont engendrées par le chevauchement. Ce problème est mentionné dans [66] et nous venons de l'illustrer dans le chapitre 2. Nous proposons ici un modèle de nucléotides avec l'influence d'un contexte unilatéral, et plus précisément avec la dépendance d'un site à son voisin de gauche. Ce contexte permet de tenir compte des phénomènes les plus influents, comme l'effet CpG ou l'effet TpA présentés au paragraphe 1.3.2 du chapitre 1. D'autre part, par souci de simplicité, nous discrétisons le modèle en une chaîne de Markov, et autorisons plusieurs substitutions par intervalle de temps.

Ces raisons nous amènent ainsi à définir le modèle décrit dans la section suivante. Nous montrons qu'il peut être vu comme une HMM, et procédons à l'estimation de ses paramètres à l'aide de l'algorithme EM pour chaînes de Markov cachées. Enfin, nous donnons les résultats de simulations réalisées pour évaluer la méthode d'inférence, et appliquons le procédé d'estimation à des séquences biologiques.

3.1 Description du modèle

Nous présentons un modèle de nucléotides avec dépendance au voisin de gauche. Le choix du côté gauche est arbitraire, et le modèle peut être vu avec une dépendance à droite sur le brin complémentaire.

3.1.1 Graphe des dépendances du modèle

Par souci de simplicité, nous définissons un modèle à temps discret qui fait évoluer une séquence initiale :

$$\mathbf{X}(1) = (X_1(1), \ldots, X_n(1))$$

vers une séquence finale :

$$\mathbf{X}(M) = (X_1(M), \ldots, X_n(M)),$$

en $M-1$ étapes évolutives, pour $M \geqslant 2$ fixé. Ces deux séquences $\mathbf{X}(1)$ et $\mathbf{X}(M)$, supposées connues, représentent respectivement un ancêtre et son descendant. Nous supposons qu'elles sont de même longueur n et à valeurs dans $\mathcal{A}^n$ avec $\mathcal{A} = \{A, C, G, T\}$, et autorisons les substitutions multiples lors des étapes évolutives. La trajectoire complète

$$(\mathbf{X}(1), \cdots, \mathbf{X}(M))$$

de $\mathbf{X} = (\mathbf{X}_i)_{i\in\{1,\ldots,n\}}$ en $M-1$ étapes évolutives est inconnue :

$$\begin{array}{cccc|l} X_1(1) & X_2(1) & \ldots & X_n(1) & \\ X_1(2) & X_2(2) & \ldots & X_n(2) & \text{temps} \\ \vdots & \vdots & \ddots & \vdots & \\ X_1(M) & X_2(M) & \ldots & X_n(M) & \downarrow \end{array}$$

Nous définissons les dépendances telles que l'état futur d'un site ne dépende que de son présent et de celui de son voisin de gauche. Cette hypothèse temporelle de **dépendance markovienne à gauche** donne son nom au modèle et est représentée par le DAG de la figure 3.1.

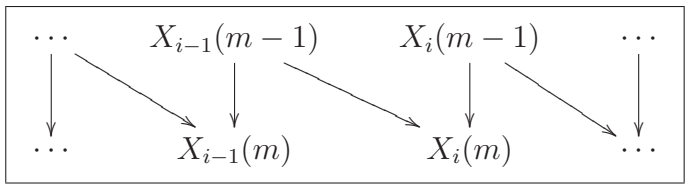


Figure 3.1 – DAG du modèle de dépendance au voisin de gauche.

Ce graphe traduit la propriété markovienne des lois conditionnelles pour tout i de $\{2,\ \ldots,\ n\}$ et tout m de $\{2,\ \ldots,\ M\}$:

$$\mathbb{P}(X_i(m) \mid \{X_j(k)\}_{j\leqslant i,\ k\leqslant m-1}) = \mathbb{P}(X_i(m) \mid \{X_j(m-1)\}_{j\leqslant i}),$$

et la dépendance à gauche réduit cette expression à :

$$\mathbb{P}(X_i(m) \mid \{X_j(m-1)\}_{j\leqslant i}) = \mathbb{P}(X_i(m) \mid X_i(m-1),\ X_{i-1}(m-1)).$$

Le modèle est supposé homogène en temps, ainsi la probabilité conditionnelle :

$$\mathbb{P}\left(X_i(m) \mid X_i(m-1),\ X_{i-1}(m-1)\right)$$

ne dépend pas de m. Il est aussi supposé uniforme et par conséquent cette probabilité ne dépend pas non plus de i.

De ce modèle « espace-temps » se déduit un modèle spatial sur les colonnes X_i qui forment ainsi une chaîne de Markov homogène :

$$X_1 \longrightarrow \cdots \longrightarrow X_i \longrightarrow \cdots \longrightarrow X_n.$$

Ces deux modèles spatiaux et temporels se superposent pour régir l'évolution de la séquence $\mathbf{X}$. La structure asymétrique des dépendances rend le modèle non réversible. Nous expliquons dans le paragraphe suivant comment ramener ce modèle dans un cadre connu pour lequel nous disposons de méthodes efficaces.

3.1.2 Notre modèle vu comme une chaîne de Markov cachée

Cette superposition de deux processus de Markov peut être traitée comme une chaîne de Markov cachée (HMM pour Hidden Markov Model). Une HMM est formée de deux processus : d'une part le processus caché non observé X_i qui forme une chaîne de Markov homogène, et d'autre part les états observés Y_i, indépendants conditionnellement aux états cachés. Les HMM, encore appelées « probabilistic functions of finite Markov chains », sont introduites en 1966 par Baum et Petrie dans [9], puis reprises dans [10, 67]. Une très bonne introduction aux HMM est donnée dans [69]. Ces modèles sont rapidement devenus populaires dans de nombreux domaines, comme la biologie, voir par exemple [22, 62], la linguistique dans [69], l'imagerie etc. Ils sont utilisés pour prendre en compte l'hétérogénéité des données au travers de différents régimes déterminés par des variables cachées, mais aussi la dépendance dans la succession des régimes, traduite par un processus markovien sur ces états cachés.

Dans le cas de notre modèle avec dépendance au voisin de gauche, la chaîne de Markov cachée est représentée par le DAG de la figure 3.2.

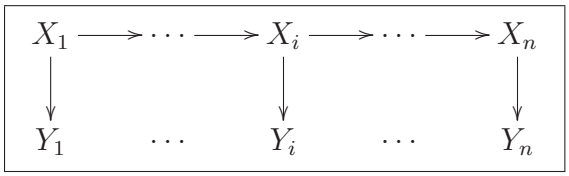


FIGURE 3.2 – DAG de la chaîne de Markov cachée.

Les états cachés sont les colonnes entières de l'évolution temporelle du site i à valeurs dans $\mathcal{A}^M$:

$$\{X_i\}_{1 \leqslant i \leqslant n}, \quad X_i = (X_i(m))_{1 \leqslant m \leqslant M},$$

et les états observés sont les états initiaux et finaux de l'évolution :

$$Y_i = (X_i(1), X_i(M)) \in \mathcal{A}^2, \quad i \in \{1, \ldots, n\}.$$

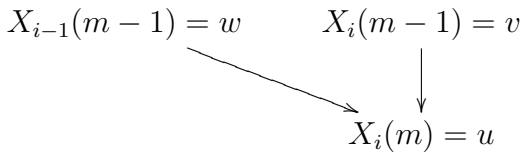
Ceux-ci sont indépendants conditionnellement aux états cachés. Notons que les observations sont même déterministes conditionnellement aux états cachés.

3.1.3 Paramètres du modèle

Détaillons à présent les paramètres utilisés dans ce modèle. Nous notons pour tous u, v et w dans $\mathcal{A}$:

$$c(u; v, w) = \mathbb{P}(X_i(m) = u | X_i(m-1) = v, X_{i-1}(m-1) = w),\ \forall m \in \{2, \dots, M\},$$

la probabilité de substitution de v par u avec le contexte (w, v), comme décrit par le schéma suivant :



Considérons l'ensemble des paramètres :

$$\mathscr{C} = \{c(u; v, w) \mid \forall (u, v, w) \in \mathcal{A}^3,\ 0 \leqslant c(u; v, w) \leqslant 1,\ \sum_{u \in \mathcal{A}} c(u; v, w) = 1\}.$$

De façon plus générale, plaçons-nous dans le cadre des HMM. Les états cachés X_i sont générés par une chaîne de Markov homogène d'ordre 1 de matrice de transition :

$$D = (D(U, V))_{U,V \in \mathcal{A}^M}, \quad D(U, V) = \mathbb{P}(X_i = V | X_{i-1} = U), \forall i \in \{2, \dots, n\}.$$

La théorie des DAG, et plus particulièrement la Propriété 2 énoncée à la section 2.1 du chapitre 2, nous permet de relier les paramètres c et D au travers des éléments des colonnes $U = (u_1, \dots, u_M)$ et $V = (v_1, \dots, v_M)$:

$$D(U, V) = \pi_{v_1} \prod_{m=2}^{M} c(v_m; v_{m-1}, u_{m-1}). \tag{3.1.1}$$

Les fréquences des nucléotides $\pi_v,\ v \in \mathcal{A}$, sont estimées par les fréquences observées une fois pour toutes pour chaque jeu de données $(\mathbf{X}(1), \mathbf{X}(M))$. La loi d'émission B de Y_i conditionnellement à l'état caché X_i est déterministe car les états observés font partie des colonnes d'états cachés pour tout i de $\{1, \dots, n\}$:

$$B = (B(U, y),\ U = (u_1, \dots, u_M) \in \mathcal{A}^M,\ y \in \mathcal{A}^2),$$

où

$$B(U, y) = \mathbb{P}(Y_i = y | X_i = U) = \mathbb{1}_{(u_1, u_M) = y}.$$

Le processus évolutif qui construit les trajectoires est initialisé par la génération de la trajectoire du premier site X_1 avec une loi multinomiale de paramètres $\pi_v,\ v \in \mathcal{A}$. Cette trajectoire constitue la première colonne du schéma des trajectoires complètes des sites. La distribution de la première colonne est notée plus généralement :

$$A = (A(U) = \mathbb{P}(X_1 = U),\ U \in \mathcal{A}^M).$$

Comme les paramètres $c(u; v, w)$ sont nombreux, nous supposons qu'ils sont des fonctions d'un paramètre multidimensionnel sous-jacent θ d'un ensemble Θ, que nous spécifierons dans la section 3.3. Ce paramètre θ nous permet de caractériser le processus biologique de substitution et constitue notre paramètre d'intérêt, c'est pourquoi nous notons à présent les paramètres des processus $c_\theta(u; v, w)$ et $D_\theta(U, V)$. Nous souhaitons estimer le vrai paramètre θ_{V} par la méthode du maximum de vraisemblance. Une HMM étant un modèle à données incomplètes, nous notons $L^o_Y(\theta) = \mathbb{P}_\theta(Y)$ la vraisemblance observée et $L^c_{Y,X}(\theta) = \mathbb{P}_\theta(Y, X)$ la vraisemblance complète, avec $\theta \in \Theta$. Les log-vraisemblances correspondantes sont notées $\ell^o_Y(\theta)$ et $\ell^c_{Y,X}(\theta)$. Comme $\mathbf{X}(1)$ est considéré comme l'ancêtre de $\mathbf{X}(M)$, nous écrivons la vraisemblance observée suivante pour les extrémités $Y_i = (X_i(1), X_i(M))$ de la trajectoire de $\mathbf{X}$, pour i variant de 1 à n :

$$\begin{aligned}
L^o_Y(\theta) &= \mathbb{P}_\theta(Y_1, \ldots, Y_n) \\
&= \sum_{x_1 \in \mathcal{A}^M} \cdots \sum_{x_n \in \mathcal{A}^M} \mathbb{P}_\theta(Y_1, \ldots, Y_n, X_1 = x_1, \ldots, X_n = x_n), \\
&= \sum_{x_1 \in \mathcal{A}^M} \cdots \sum_{x_n \in \mathcal{A}^M} A(x_1) \prod_{i=1}^{n} B(x_i, Y_i) \prod_{i=2}^{n} D_\theta(x_{i-1}, x_i), \\
&= \sum_{x_1 \in \mathcal{A}^M} \cdots \sum_{x_n \in \mathcal{A}^M} L^c_{Y,x}(\theta).
\end{aligned}$$

La somme sur tous les états cachés possibles devient rapidement impossible à calculer lorsque n et M augmentent. Ainsi, l'estimation de θ_{V} requiert une procédure adaptée présentée ci-dessous.

3.2 Estimation des paramètres

Pour obtenir l'estimateur $\hat{\theta}_n$ du maximum de vraisemblance de θ_{V} pour n fixé, il faut maximiser la vraisemblance des observations en θ. L'écriture précédente de $L^o_\theta(Y)$ montre à quel point cette opération est difficile. De plus, la maximisation d'une telle somme de termes ne peut s'effectuer de façon simple, et demande une résolution fastidieuse parce que l'annulation de la dérivée ne donne pas d'expression explicite pour $\hat{\theta}_n$. La procédure qui suit maximise donc en lieu et place l'espérance conditionnelle aux observations de la vraisemblance des données complètes

$L^c_{Y,X}(\theta)$ qui est plus aisée à manipuler. Nous utilisons ainsi une version de **l'algorithme Expectation-Maximization** (EM) développé par Dempster, Laird et Rubin dans [27], spécifique aux chaînes de Markov cachées (**EM-HMM**). Cet algorithme, aussi appelé **algorithme de Baum-Welch**, est une procédure d'estimation générale qui permet d'approcher par itérations successives l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ_{V} par une suite $(\theta_n^{(k)})_{k\geqslant 0}$ en augmentant à chaque étape la vraisemblance $L^o_Y(\theta)$ à travers $L^c_{Y,X}(\theta)$.

La log-vraisemblance complète est :

$$\ell^c_{Y,X}(\theta) = \ln A(X_1) + \sum_{i=1}^{n} \ln B(X_i, Y_i) + \sum_{i=2}^{n} \sum_{U\in\mathcal{A}^M} \sum_{V\in\mathcal{A}^M} \mathbb{1}_{X_{i-1}=U, X_i=V} \ln D_\theta(U,V).$$

Soit Q l'espérance de $\ell^c_{Y,X}(\theta)$ conditionnellement aux observations et pour le paramètre $\theta_n^{(k)}$, obtenu après k itérations :

$$\begin{aligned} Q(\theta, \theta_n^{(k)}) &= \mathbb{E}_{\theta_n^{(k)}}(\ell^c_{Y,X}(\theta)|Y), \\ &= \Gamma + \sum_{i=2}^{n} \sum_{U\in\mathcal{A}^M} \sum_{V\in\mathcal{A}^M} \mathbb{P}_{\theta_n^{(k)}}(X_{i-1}=U, X_i=V|Y) \ln D_\theta(U,V), \end{aligned}$$

où Γ est une constante indépendante de θ.

Voici le déroulement de l'algorithme. Commençons par choisir une valeur d'initialisation $\theta_n^{(0)}$ du paramètre. Une boucle EM est constituée de deux étapes :

- **Étape E** : calculer $Q(\theta, \theta_n^{(k)})$,
- **Étape M** : poser $\theta_n^{(k+1)} = \arg\max_\theta Q(\theta, \theta_n^{(k)})$.

Cet algorithme augmente au fil des boucles EM la vraisemblance des observations. En effet, soit $\mathbb{P}_\theta(X|Y)$ la loi marginale de X conditionnellement à Y pour le paramètre θ. La vraisemblance L^c peut se décomposer en :

$$L^c_{Y,X}(\theta) = \mathbb{P}_\theta(X|Y)\, L^o_Y(\theta),$$

ce qui donne pour les log-vraisemblances l'égalité suivante pour tout θ :

$$\ell^c_{Y,X}(\theta) = \ln \mathbb{P}_\theta(X|Y) + \ell^o_Y(\theta).$$

Ainsi :

$$\begin{aligned} Q(\theta, \theta_n^{(k)}) &= \mathbb{E}_{\theta_n^{(k)}}(\ell^c_{Y,X}(\theta)|Y), \\ &= \mathbb{E}_{\theta_n^{(k)}}(\ln \mathbb{P}_\theta(X|Y)|Y) + \mathbb{E}_{\theta_n^{(k)}}(\ell^o_Y(\theta)|Y). \end{aligned}$$

Posons

$$H(\theta, \theta_n^{(k)}) = \mathbb{E}_{\theta_n^{(k)}}(\ln \mathbb{P}_\theta(X|Y)|Y).$$

Alors :

$$\begin{aligned} Q(\theta, \theta_n^{(k)}) &= H(\theta, \theta_n^{(k)}) + \mathbb{E}_{\theta_n^{(k)}}(\ell_Y^o(\theta)|Y), \\ &= H(\theta, \theta_n^{(k)}) + \ell_Y^o(\theta), \end{aligned}$$

et

$$\ell_Y^o(\theta) = Q(\theta, \theta_n^{(k)}) - H(\theta, \theta_n^{(k)}).$$

Posons $\theta_n^{(k+1)} = \arg\max_\theta Q(\theta, \theta_n^{(k)})$. Comme $\theta_n^{(k+1)}$ est la valeur qui maximise la quantité $Q(\theta, \theta_n^{(k)})$, nous en déduisons directement que :

$$Q(\theta_n^{(k+1)}, \theta_n^{(k)}) \geqslant Q(\theta_n^{(k)}, \theta_n^{(k)}).$$

De plus,

$$\begin{aligned} H(\theta_n^{(k+1)}, \theta_n^{(k)}) - H(\theta_n^{(k)}, \theta_n^{(k)}) &= \mathbb{E}_{\theta_n^{(k)}}(\ln \mathbb{P}_{\theta_n^{(k+1)}}(X|Y)|Y) - \mathbb{E}_{\theta_n^{(k)}}(\ln \mathbb{P}_{\theta_n^{(k)}}(X|Y)|Y), \\ &= \mathbb{E}_{\theta_n^{(k)}}\left(\ln\left[\frac{\mathbb{P}_{\theta_n^{(k+1)}}(X|Y)}{\mathbb{P}_{\theta_n^{(k)}}(X|Y)}\right]\Big|Y\right). \end{aligned}$$

L'inégalité de Jensen conditionnelle implique alors :

$$\begin{aligned} H(\theta_n^{(k+1)}, \theta_n^{(k)}) - H(\theta_n^{(k)}, \theta_n^{(k)}) &\leqslant \ln\left[\mathbb{E}_{\theta_n^{(k)}}\left(\frac{\mathbb{P}_{\theta_n^{(k+1)}}(X|Y)}{\mathbb{P}_{\theta_n^{(k)}}(X|Y)}\Big|Y\right)\right], \\ &= \ln\left[\sum_x \frac{\mathbb{P}_{\theta_n^{(k+1)}}(X=x|Y)}{\mathbb{P}_{\theta_n^{(k)}}(X=x|Y)}\mathbb{P}_{\theta_n^{(k)}}(X=x|Y)\right], \\ &= \ln\left[\sum_x \mathbb{P}_{\theta_n^{(k+1)}}(X=x|Y)\right] = \ln 1 = 0. \end{aligned}$$

Nous en déduisons : $H(\theta_n^{(k)}, \theta_n^{(k)}) \geqslant H(\theta_n^{(k+1)}, \theta_n^{(k)})$, ce qui nous permet de conclure que :

$$\ell_Y^o(\theta_n^{(k+1)}) \geqslant \ell_Y^o(\theta_n^{(k)}).$$

L'algorithme s'arrête soit lorsque le nombre maximum de boucles autorisées est atteint, soit lorsque l'accroissement de la log-vraisemblance devient négligeable d'une itération à l'autre, c'est-à-dire quand pour un certain $k = K$ et un ϵ_1 fixés :

$$\left|\ell_Y^o(\theta_n^{(K+1)}) - \ell_Y^o(\theta_n^{(K)})\right| < \epsilon_1.$$

En pratique, nous vérifions que pour un ϵ_2 fixé :

$$\left| \frac{Q(\theta_n^{(K+1)}, \theta_n^{(K+1)}) - Q(\theta_n^{(K)}, \theta_n^{(K)})}{Q(\theta_n^{(K)}, \theta_n^{(K)})} \right| < \epsilon_2,$$

et nous posons $\tilde{\theta}_n = \theta_n^{(K+1)}$ comme valeur approchée de $\hat{\theta}_n$.

Cette procédure garantit la convergence vers un maximum local ou un point stationnaire. Il faut ensuite soit analyser les données plus précisément pour s'assurer que le paramètre trouvé est un maximum global, soit répéter l'estimation avec plusieurs valeurs initiales différentes. Les résultats de convergence et de normalité de l'estimateur du maximum de vraisemblance justifiant l'approche par maximum de vraisemblance ont été prouvés pour la première fois par Baum et Petrie dans [9]. Les résultats sur la convergence de l'algorithme EM figurent dans [81].

3.2.1 Algorithme Forward-Backward pour le calcul de l'espérance

Soit $y_i = (x_i(1), x_i(M))$ une réalisation de Y_i. Nous notons :

$$\forall 1 \leqslant i \leqslant j \leqslant n, \quad y_i^j = (y_i, \ldots, y_j) \quad \text{et} \quad y_1^n = y = (y_1, \ldots, y_n).$$

Pour l'étape E, calculer la quantité :

$$Q(\theta, \theta_n^{(k)}) = \Gamma + \sum_{i=2}^{n} \sum_{U \in \mathcal{A}^M} \sum_{V \in \mathcal{A}^M} \mathbb{P}_{\theta_n^{(k)}}(X_{i-1} = U, X_i = V | y) \ln D_\theta(U, V)$$

nécessite de connaître les probabilités $\mathbb{P}_{\theta_n^{(k)}}(X_{i-1} = U, X_i = V|y)$ en fonction notamment des éléments $D_{\theta_n^{(k)}}(U, V)$ de la matrice de transition. Nous noterons ces quantités plus simplement $\mathbb{P}_k(X_{i-1} = U, X_i = V|y)$ et $D_k(U, V)$. Ces probabilités sont obtenues à l'aide d'une série de boucles, nommée algorithme Forward-Backward ou algorithme de descente-remontée. Les équations utilisées dans les boucles sont les équations prédictives (EP), de filtrage (EF) et de lissage (EL1) et (EL2) :

(EP) :

$$\begin{aligned}
\mathbb{P}_k(X_i = V | y_1^{i-1}) &= \sum_{U \in \mathcal{A}^M} \mathbb{P}_k(X_{i-1} = U, X_i = V | y_1^{i-1}), \\
&= \sum_{U \in \mathcal{A}^M} \mathbb{P}_k(X_i = V | X_{i-1} = U, y_1^{i-1}) \mathbb{P}_k(X_{i-1} = U | y_1^{i-1}), \\
&= \sum_{U \in \mathcal{A}^M} \mathbb{P}_k(X_i = V | X_{i-1} = U) \mathbb{1}_{(u_1, u_M) = y_{i-1}} \mathbb{P}_k(X_{i-1} = U | y_1^{i-1}), \\
&= \sum_{U \in \mathcal{A}^M} D_k(U, V) \mathbb{P}_k(X_{i-1} = U | y_1^{i-1}) \mathbb{1}_{(u_1, u_M) = y_{i-1}}.
\end{aligned}$$

(EF) :

$$\begin{aligned}
\mathbb{P}_k(X_i = U|y_1^i) &= \mathbb{P}_k(X_i = U|y_1^{i-1}, y_i), \\
&= \frac{\mathbb{P}_k(X_i = U, y_i|y_1^{i-1})}{\mathbb{P}_k(y_i|y_1^{i-1})}, \\
&= \frac{\mathbb{P}_k(X_i = U, y_i|y_1^{i-1})}{\sum_{V\in\mathcal{A}^M}\mathbb{P}_k(X_i = V, y_i|y_1^{i-1})}, \\
&= \frac{\mathbb{P}_k(y_i|X_i = U, y_1^{i-1})\mathbb{P}_k(X_i = U|y_1^{i-1})}{\sum_{V\in\mathcal{A}^M}\mathbb{P}_k(y_i|X_i = V, y_1^{i-1})\mathbb{P}_k(X_i = V|y_1^{i-1})}, \\
&= \frac{\mathbb{1}_{(u_1,u_M)=y_i}\mathbb{P}_k(X_i = U|y_1^{i-1})}{\sum_{V\in\mathcal{A}^M}\mathbb{1}_{(v_1,v_M)=y_i}\mathbb{P}_k(X_i = V|y_1^{i-1})}.
\end{aligned}$$

(EL1) :

$$\begin{aligned}
\mathbb{P}_k(X_{i-1} = U, X_i = V|y) &= \mathbb{P}_k(X_{i-1} = U|X_i = V, y)\mathbb{P}_k(X_i = V|y), \\
&= \mathbb{P}_k(X_{i-1} = U|X_i = V, y_1^{i-1})\mathbb{P}_k(X_i = V|y)\mathbb{1}_{(v_0,v_M)=y_i}, \\
&= \frac{\mathbb{P}_k(X_i = V, X_{i-1} = U|y_1^{i-1})\mathbb{P}_k(X_i = V|y)}{\mathbb{P}_k(X_i = V|y_1^{i-1})}\mathbb{1}_{(v_1,v_M)=y_i}, \\
&= \frac{\mathbb{P}_k(X_i = V|X_{i-1} = U, y_1^{i-1})\mathbb{P}_k(X_{i-1} = U|y_1^{i-1})}{\mathbb{P}_k(X_i = V|y_1^{i-1})} \cdot \\
&\qquad \mathbb{P}_k(X_i = V|y)\mathbb{1}_{(v_1,v_M)=y_i}, \\
&= \frac{\mathbb{P}_k(X_i = V|X_{i-1} = U)\mathbb{1}_{(u_1,u_M)=y_{i-1}}\mathbb{P}_k(X_{i-1} = U|y_1^{i-1})}{\mathbb{P}_k(X_i = V|y_1^{i-1})} \cdot \\
&\qquad \mathbb{P}_k(X_i = V|y)\mathbb{1}_{(v_1,v_M)=y_i}, \\
&= \frac{\mathbb{P}_k(X_{i-1} = U|y_1^{i-1})\mathbb{P}_k(X_i = V|y)}{\mathbb{P}_k(X_i = V|y_1^{i-1})} \cdot \\
&\qquad D_k(U,V)\mathbb{1}_{(u_1,u_M)=y_{i-1}}\mathbb{1}_{(v_1,v_M)=y_i}.
\end{aligned}$$

(EL2) :

$$\begin{aligned}
\mathbb{P}_k(X_{i-1} = U|y) &= \sum_{V\in\mathcal{A}^M}\mathbb{P}_k(X_{i-1} = U, X_i = V|y), \\
&= \mathbb{P}_k(X_{i-1} = U|y_1^{i-1})\mathbb{1}_{(u_1,u_M)=y_{i-1}} \cdot \\
&\qquad \sum_{V\in\mathcal{A}^M}\frac{\mathbb{P}_k(X_i = V|y)}{\mathbb{P}_k(X_i = V|y_1^{i-1})}D_k(U,V)\mathbb{1}_{(v_1,v_M)=y_i}.
\end{aligned}$$

L'algorithme Forward-Backward consiste à calculer la valeur de $\mathbb{P}_k(X_1 = V|y_1)$ qui est soit 0, soit 1, puis à utiliser en alternance les équations (EP) et (EF) lors de la descente jusqu'aux valeurs des $\mathbb{P}_k(X_n = U|y_1^n)$, $U \in \mathcal{A}^M$. Il faut ensuite utiliser les équations de lissage (EL1) et (EL2) lors de la remontée pour aboutir aux $\mathbb{P}_k(X_{i-1} = U, X_i = V|y)$ pour tous U et V dans $\mathcal{A}^M$. Cette procédure est représentée schématiquement en figure 3.3.

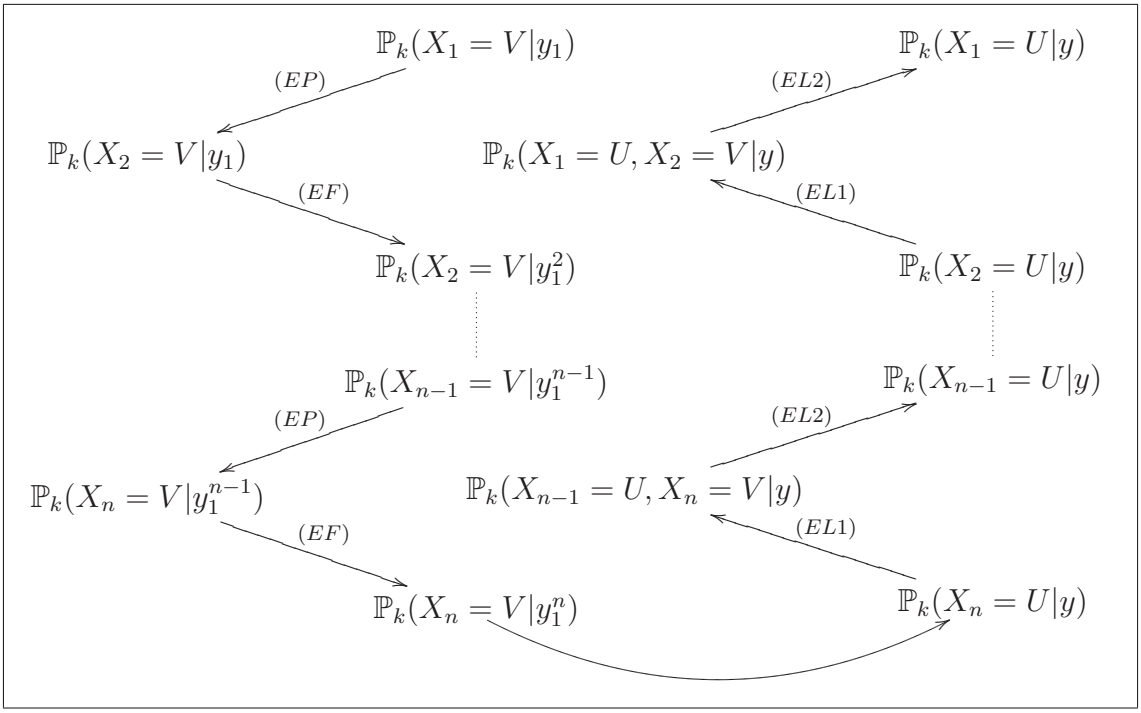


Figure 3.3 – Déroulement de l'algorithme Forward-Backward.

3.2.2 Étape de maximisation

L'étape M, c'est-à-dire la maximisation de Q, est réalisée numériquement. Nous verrons plus bas que notre choix de paramètres facilite ce calcul.

3.3 Résultats

3.3.1 Paramètres

Les paramètres $c_\theta(u; v, w)$ apparaissant dans l'écriture (3.1.1) des coefficients de la matrice de transition D, au nombre de 48 soit 64 composantes soumises à 16 contraintes de sommation à 1, sont trop généraux pour pouvoir être interprétés d'un point de vue biologique. En effet, comme le montre [7], certaines dépendances sont plus pertinentes que d'autres, par exemple l'effet CpG qui est le principal effet de voisinage observé. De plus, ils sont trop nombreux pour être estimés en un temps raisonnable.

C'est pourquoi nous avons choisi un paramètre intermédiaire de taille trois permettant de décrire plus précisément les événements évolutifs. Ce paramètre s'écrit $\theta = (\mu, \tau, \gamma) \in [0, 1]^3$, avec :

- μ la probabilité pour un site de subir une substitution,
- τ un facteur multiplicatif prenant en compte la différence entre transitions et transversions,
- γ un facteur multiplicatif prenant en compte la perte d'un dinucléotide CpG ou TpA durant la substitution.

Pour $u \in \mathcal{A}$, soit u^* le nucléotide complémentaire de u dans sa classe chimique. Par exemple, si u est la purine A, alors $u \in \mathcal{R}$ et $u^* = G$. Nous définissons l'indicatrice de Watson-Kronecker ϵ comme étant la fonction suivante :

$$\forall (u, v) \in \mathcal{A}^2, \quad \epsilon(u, v) = \mathbb{1}_{v=u^*}.$$

Nous notons également κ l'indicatrice d'un contexte différent de CpG et de TpA :

$$\forall (w, v) \in \mathcal{A}^2, \quad \kappa(w, v) = \mathbb{1}_{(w,v)\neq(C,G)} \cdot \mathbb{1}_{(w,v)\neq(T,A)}.$$

Nous définissons alors les probabilités avec dépendance au contexte $c_\theta(u; v, w)$ par :

$$c_\theta(u; v, w) = \begin{cases} \mu \cdot \tau^{\epsilon(u,v)} \cdot \left(\dfrac{1-\tau}{2}\right)^{1-\epsilon(u,v)} \cdot \gamma^{\kappa(w,v)} & \text{si } v \neq u, \\ 1 - \mu \cdot \gamma^{\kappa(w,v)} & \text{si } v = u. \end{cases}$$

Ainsi, lorsque $u \neq v$, la probabilité de substitution est proportionnelle :

- à la probabilité globale de substitution μ,
- au facteur τ si la substitution est une transition, et au facteur $(1 - \tau)/2$ si la substitution est une transversion,
- au facteur γ si le contexte n'est ni CpG ni TpA.

Le facteur $1/2$ apparaît dans le cas des transversions parce qu'il existe deux fois plus de types de transversions que de types de transitions. En effet, les transitions

sont les substitutions $A \to G,\ G \to A,\ C \to T,\ T \to C$, alors que les transversions possibles sont $A \to C,\ C \to A,\ A \to T,\ T \to A,\ G \to C,\ C \to G,\ G \to T,\ T \to G$. Ainsi, pour chaque contexte (v, w) de $\mathcal{A}^2$, $\sum_{u \in \mathcal{A}} c_\theta(u; v, w) = 1$. Voici une autre présentation possible pour les probabilités $\mathrm{c}_\theta(u; v, w)$:

- dans un contexte CpG/TpA, c'est-à-dire $(w, v) = (C, G)$ ou (T, A) :

$$c_\theta(u; v, w) = \begin{cases} 1 - \mu & \text{si } v = u, \\ \mu \cdot \tau & \text{si } v = u^*, \\ \mu \cdot \left(\dfrac{1 - \tau}{2} \right) & \text{sinon,} \end{cases}$$

- dans tout autre type de contexte :

$$c_\theta(u; v, w) = \begin{cases} 1 - \mu \cdot \gamma & \text{si } v = u, \\ \mu \cdot \tau \cdot \gamma & \text{si } v = u^*, \\ \mu \cdot \left(\dfrac{1 - \tau}{2} \right) \cdot \gamma & \text{sinon.} \end{cases}$$

Notons que le paramètre θ appartient au compact $[0, 1]^3$, ce qui facilite la mise en œuvre numérique de l'étape de maximisation.

3.3.2 Identifiabilité du modèle

Pour procéder à l'estimation de c_θ à travers l'estimation de θ, il faudrait nous assurer que le modèle que nous considérons est identifiable, c'est-à-dire que pour tout $n \in \mathbb{N}^*$, tout $M \geqslant 2$, tout couple de séquences $(\mathbf{X}(1), \mathbf{X}(M))$ de longueur n et tout couple de paramètres $\theta \neq \theta'$, nous avons :

$$\mathbb{P}_\theta(\mathbf{X}(1), \mathbf{X}(M)) \neq \mathbb{P}_{\theta'}(\mathbf{X}(1), \mathbf{X}(M)).$$

Il est connu que les modèles de chaînes de Markov cachées ne sont pas identifiables au sens strict. En effet, la vraisemblance $L_Y^o(\theta) = \mathbb{P}_\theta(\mathbf{X}(1), \mathbf{X}(M))$ donnée au paragraphe 3.1.3 est inchangée par une permutation des indices des états cachés. Nous regardons donc l'identifiabilité à permutation près. Voici le diagramme faisant intervenir les différents paramètres utilisés :

$$\begin{array}{ccc} \theta = (\mu, \tau, \gamma) & \longrightarrow & \mathbb{P}_\theta(\mathbf{X}(1), \mathbf{X}(M)) \\ \Big\downarrow \mathbf{1} & & \Big\uparrow \mathbf{3} \\ \mathscr{C}_\theta = \{c_\theta(u; v, w)\}_{(u,v,w) \in \mathcal{A}^3} & \xrightarrow{-\frac{1}{2}} & D_\theta = (D_\theta(U, V))_{(U,V) \in (\mathcal{A}^M)^2} \end{array}$$

Selon ce diagramme, l'identifiabilité peut être démontrée de deux manières différentes. La première est symbolisée par la flèche directe. L'injectivité de cette application est difficile à démontrer au vu de l'expression des $\mathbb{P}_\theta(\mathbf{X}(1), \mathbf{X}(M))$. La seconde suit les trois flèches pointillées.

L'injectivité de la flèche **1** est immédiate à partir de l'écriture des taux. En effet, considérons deux paramètres $\theta = (\mu, \tau, \gamma)$ et $\theta' = (\mu', \tau', \gamma')$. Supposons que :

$$c_\theta(u; v, w) = c_{\theta'}(u; v, w), \quad \forall (u, v, w) \in \mathcal{A}^3,$$

et plaçons-nous pour commencer dans un contexte CpG. L'expression des taux nous donne $1 - \mu = 1 - \mu'$, et $\mu \cdot \tau = \mu' \cdot \tau'$. Nous en déduisons $\mu = \mu'$ puis $\tau = \tau'$ car les paramètres sont strictement positifs. Hors d'un contexte CpG/TpA, l'égalité $\mu \cdot \tau \cdot \gamma = \mu' \cdot \tau' \cdot \gamma' = \mu \cdot \tau \cdot \gamma'$ nous permet alors d'écrire $\gamma = \gamma'$, et de conclure que $\theta = \theta'$.

Considérons à présent la flèche **2**, et raisonnons sur les colonnes. Supposons que pour deux paramètres θ et θ' :

$$D_\theta(U, V) = D_{\theta'}(U, V), \quad \forall (U, V) \in \left(\mathcal{A}^M\right)^2. \tag{3.3.1}$$

Choisissons les colonnes $U = (u_1, \ldots, u_M)$, $V^1 = (v_1^1, \ldots, v_M^1)$ et $V^2 = (v_1^2, \ldots, v_M^2)$ telles que V^1 et V^2 soient identiques à l'exception de leur dernier élément $v_M^1 \neq v_M^2$. Nous calculons le rapport entre les coefficients $D_\theta(U, V^1)$ et $D_\theta(U, V^2)$, puis utilisons l'hypothèse (3.3.1) pour écrire :

$$\frac{D_\theta(U, V^1)}{D_\theta(U, V^2)} = \frac{D_{\theta'}(U, V^1)}{D_{\theta'}(U, V^2)}.$$

La formule (3.1.1) appliquée à D_θ et à $D_{\theta'}$ donne :

$$\frac{\pi_{v_1^1} \prod_{m=2}^M c_\theta(v_m^1; v_{m-1}^1, u_{m-1})}{\pi_{v_1^2} \prod_{m=2}^M c_\theta(v_m^2; v_{m-1}^2, u_{m-1})} = \frac{\pi_{v_1^1} \prod_{m=2}^M c_{\theta'}(v_m^1; v_{m-1}^1, u_{m-1})}{\pi_{v_1^2} \prod_{m=2}^M c_{\theta'}(v_m^2; v_{m-1}^2, u_{m-1})}.$$

La similarité des colonnes a pour conséquence que ces produits se réduisent tous au terme correspondant à $m = M$, impliquant ainsi :

$$\frac{c_\theta(v_M^1; v_{M-1}^1, u_{M-1})}{c_\theta(v_M^2; v_{M-1}^2, u_{M-1})} = \frac{c_{\theta'}(v_M^1; v_{M-1}^1, u_{M-1})}{c_{\theta'}(v_M^2; v_{M-1}^2, u_{M-1})}.$$

Nous sommons ensuite les deux membres sur toutes les valeurs possibles de v_M^1, ce qui conduit à l'égalité suivante :

$$\frac{1}{c_\theta(v_M^2; v_{M-1}^2, u_{M-1})} = \frac{1}{c_{\theta'}(v_M^2; v_{M-1}^2, u_{M-1})},$$

puisque

$$\sum_{v_M^1 \in \mathcal{A}} c_\theta(v_M^1; v_{M-1}^1, u_{M-1}) = \sum_{v_M^1 \in \mathcal{A}} c_{\theta'}(v_M^1; v_{M-1}^1, u_{M-1}) = 1.$$

Ceci nous donne enfin $c_\theta(u; v, w) = c_{\theta'}(u; v, w)$ pour tous u, v et w dans $\mathcal{A}$, et donc l'injectivité de la flèche 2.

L'injectivité de la flèche **3** nous amène à considérer l'identifiabilité des HMM probabilistes traitée par Petrie dans [67]. Le résultat de Petrie ne s'applique cependant pas dans notre cas, car nous considérons en réalité une fonction déterministe d'une chaîne de Markov. En effet, les hypothèses du théorème concluant à l'identifiabilité incluent la non nullité de $\prod_{U \in \mathcal{A}^M} \prod_{y \in \mathcal{A}^2} B(U, y)$, et cette condition n'est pas vérifiée car pour tous $U \in \mathcal{A}^M$ et $y \in \mathcal{A}^2$, $B(U, y)$ vaut $\mathbb{1}_{(u_1, u_M)=y}$. Un résultat sur l'identifiabilité des fonctions déterministes de chaînes de Markov est donné par le théorème 2 de l'article [37] de Gilbert qui décrit l'ensemble des matrices équivalentes à la matrice de transition D_θ. Malheureusement, le décompte des paramètres identifiables à l'aide de la formule donnée en corollaire du théorème s'avère très difficile. Ainsi, nous ne pouvons conclure par ce biais à l'identifiabilité du modèle.

3.3.3 Simulations

Protocole La procédure d'estimation est implémentée dans le langage R à l'aide du package seqinr ([17, 68]). Nous illustrons les performances de l'algorithme sur des séquences simulées dans différentes situations, en faisant varier leur longueur n, le nombre d'étapes évolutives $M - 1$, et la valeur du paramètre θ. Les simulations sont conçues comme suit.

Nous supposons que les fréquences des nucléotides valent toutes $1/4$. La valeur M varie de 3 à 5, cette dernière valeur étant empiriquement la plus grande permettant un calcul d'une durée raisonnable. En effet, la complexité algorithmique d'une étape E de l'algorithme EM est linéaire en n et exponentielle en M. Rappelons que « pb » désigne l'unité de longueur des séquences en paires de bases. Les séquences simulées ont des longueurs de $n = 100$, 500, 1 000, 2 000, 3 000, 4 000 et 5 000 pb. Les différentes valeurs étudiées du paramètre θ figurent dans la table 3.1.

Justifions le choix de ces valeurs. Pour toute valeur de μ, une situation dite **d'évolution neutre** apparaît lorsque le processus ne tient compte ni de la classe du site subissant la substitution, ni de son contexte.

- Pour le deuxième paramètre τ, l'évolution est neutre lorsque les facteurs multiplicatifs τ et $(1 - \tau)/2$ sont égaux. L'égalité a donc lieu pour $\tau = 1/3$, et le modèle ne différencie pas les transitions des transversions. Les transitions sont favorisées lorsque $\tau > 1/3$, et leur fréquence augmente quand τ tend vers 1. Par exemple, si nous voulons que les transitions soient huit fois plus fréquentes que les transversions, il faut prendre τ tel que $\tau = 8(1 - \tau)/2$, soit $\tau = 0.8$.

θ	μ	τ	γ
θ_1	0.01	$1/3$	1
θ_2	0.1	$1/3$	1
θ_3	0.1	0.5	1
θ_4	0.1	0.8	1
θ_5	0.1	$1/3$	$1/3$
θ_6	0.1	$1/3$	0.1
θ_7	0.01	0.8	$1/3$

TABLE 3.1 – Paramètres utilisés lors des simulations.

- Le cas neutre pour le troisième paramètre apparaît lorsque le modèle ne tient pas compte du contexte, ce qui est le cas pour $\gamma = 1$. Pour favoriser les mutations à partir d'un contexte CpG/TpA, comme dans les séquences étudiées dans [16], γ doit être petit, c'est-à-dire proche de 0. Si nous voulons par exemple que de telles substitutions soient dix fois plus fréquentes que les mutations sans contexte CpG/TpA, nous choisissons $\gamma = 0.1$.

Ainsi, un modèle ayant pour paramètre $\theta = (\mu, 1/3, 1)$ avec $\mu \in]0, 1]$ engendre une évolution neutre. Dans ce cas, nous pouvons rapprocher notre modèle du modèle de nucléotides de [47] qui possède un seul paramètre. Si τ est quelconque et γ vaut 1, notre modèle peut être comparé au modèle de Kimura dans [50]. Des comparaisons avec ces deux modèles sont effectuées plus bas.

Pour chaque triplet de valeurs $(n, M, \theta_{\mathrm{V}})$, où $\theta_{\mathrm{V}} = (\mu_{\mathrm{V}}, \tau_{\mathrm{V}}, \gamma_{\mathrm{V}})$ désigne la « vraie » valeur du paramètre que nous voulons estimer, nous générons des répétitions de la procédure suivante, respectivement 50 répétitions pour $M \in \{3, 4\}$ et 25 répétitions pour $M = 5$, afin de construire des boîtes à moustaches pour chaque coordonnée de θ_{V}. La procédure en question comprend deux étapes. La première étape consiste à générer deux séquences reliées par le processus d'évolution de paramètre θ_{V}. Pour cela, nous simulons la première séquence $\mathbf{X}(1)$ et la trajectoire du premier site par des lois multinomiales équiréparties dans $\mathcal{A}$. Nous faisons ensuite évoluer $\mathbf{X}(1)$ en $M-1$ étapes pour compléter toutes les trajectoires, jusqu'à obtenir $\mathbf{X}(M)$. La séquence $\mathbf{X}(M)$ est donc bien un descendant de la séquence $\mathbf{X}(1)$. Dans la seconde étape, l'algorithme EM estime θ_{V} par une valeur notée θ_{EM} à partir des séquences $\mathbf{X}(1)$ et $\mathbf{X}(M)$.

Nous avons empiriquement constaté que la valeur $\theta_{\mathrm{init}} = (0.5, 0.5, 0.5)$ du paramètre initial nécessaire à la procédure numérique de maximisation donne de meilleurs résultats qu'une valeur initiale choisie au hasard dans $[0, 1]^3$. En effet, des valeurs comme $\theta_{\mathrm{init}} = (0.9, 0.7, 0.8)$ conduisent souvent l'algorithme à un point de la frontière ayant une ou plusieurs coordonnées égales à 0 ou 1.

Les résultats sont donnés par les graphes 3.4 à 3.13.

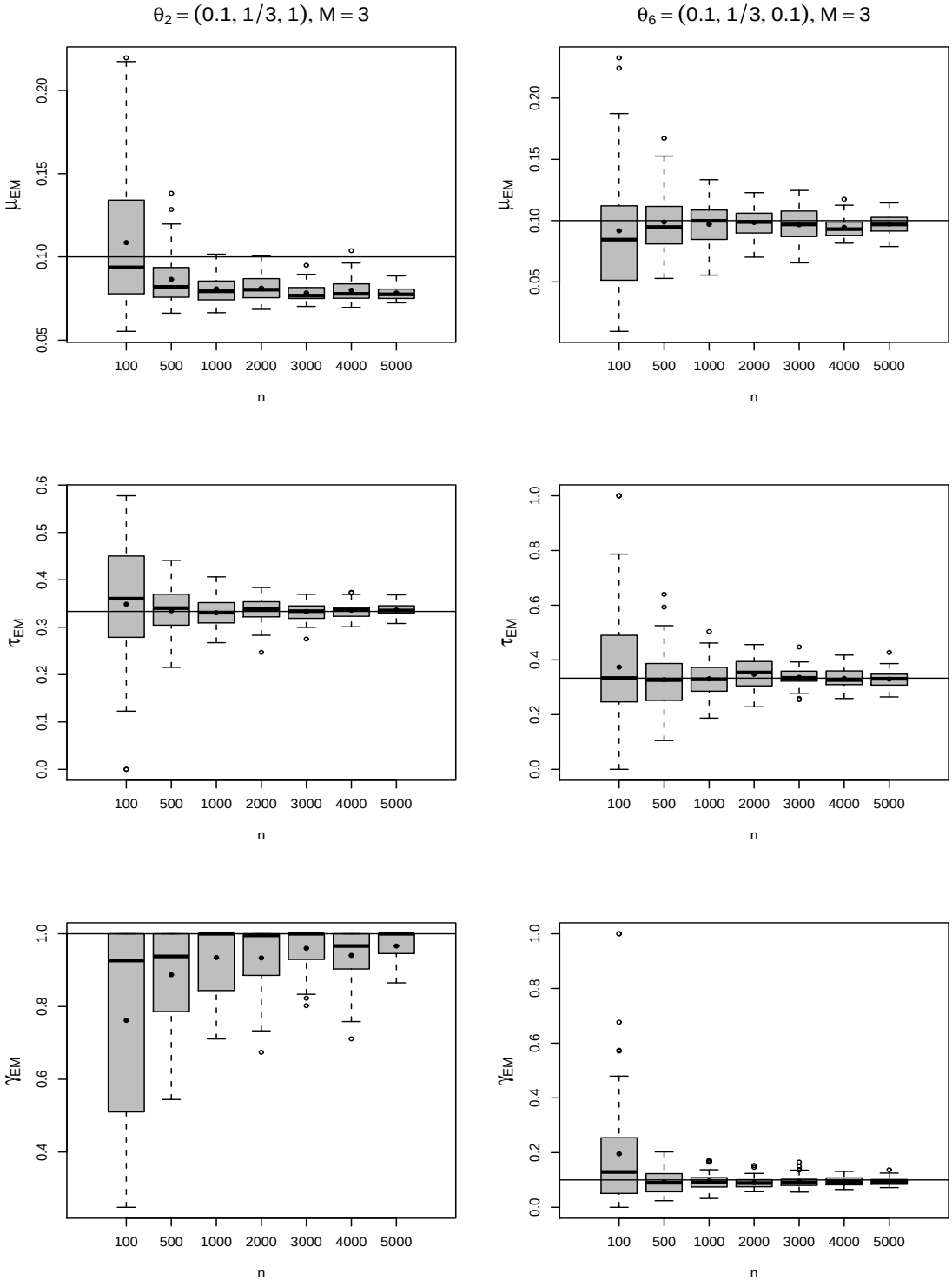


Figure 3.4 – Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_2 et θ_6 pour $M = 3$ et pour des valeurs croissantes de n.

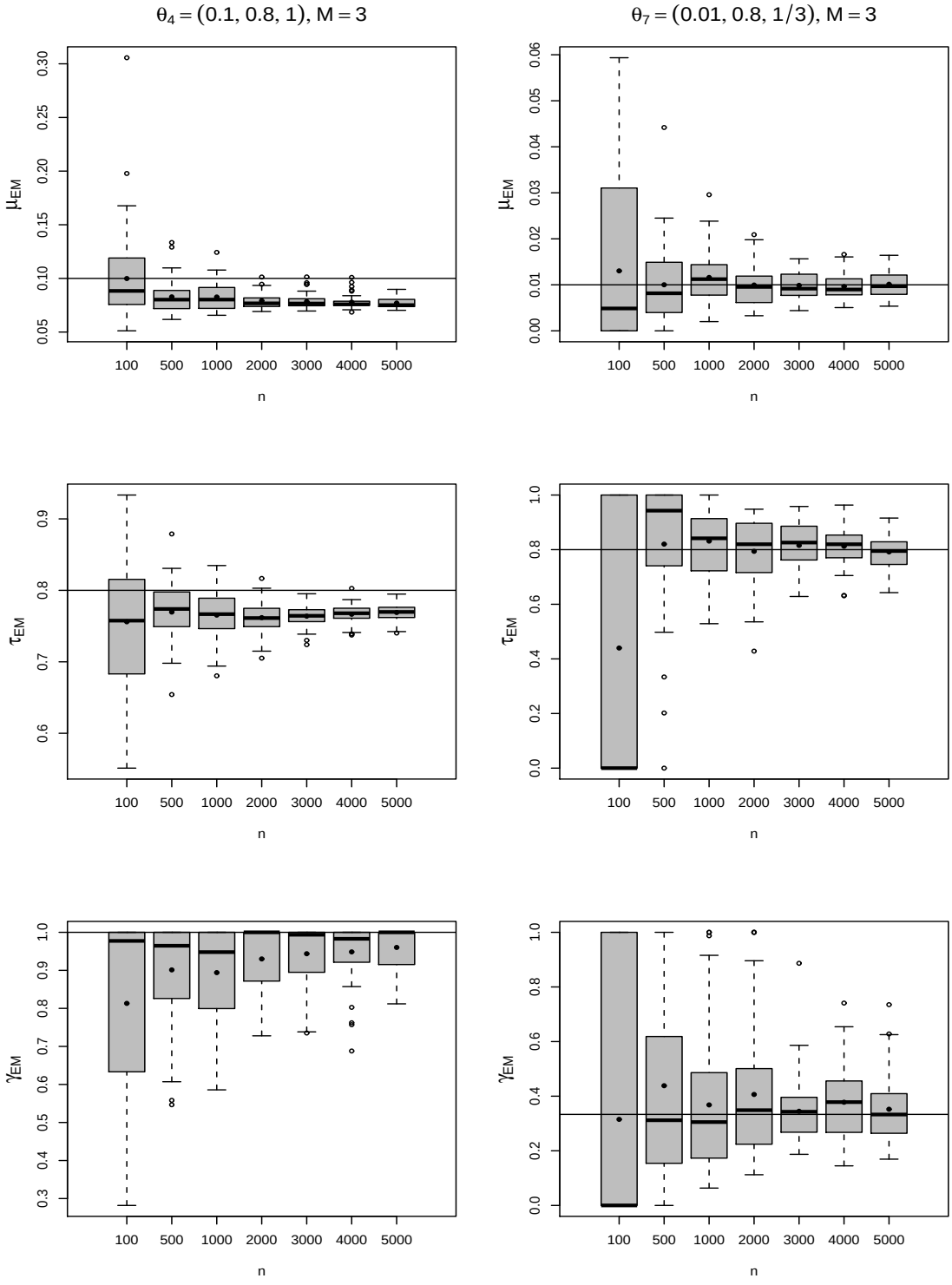


FIGURE 3.5 – Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_4 et θ_7 pour $M = 3$ et pour des valeurs croissantes de n.

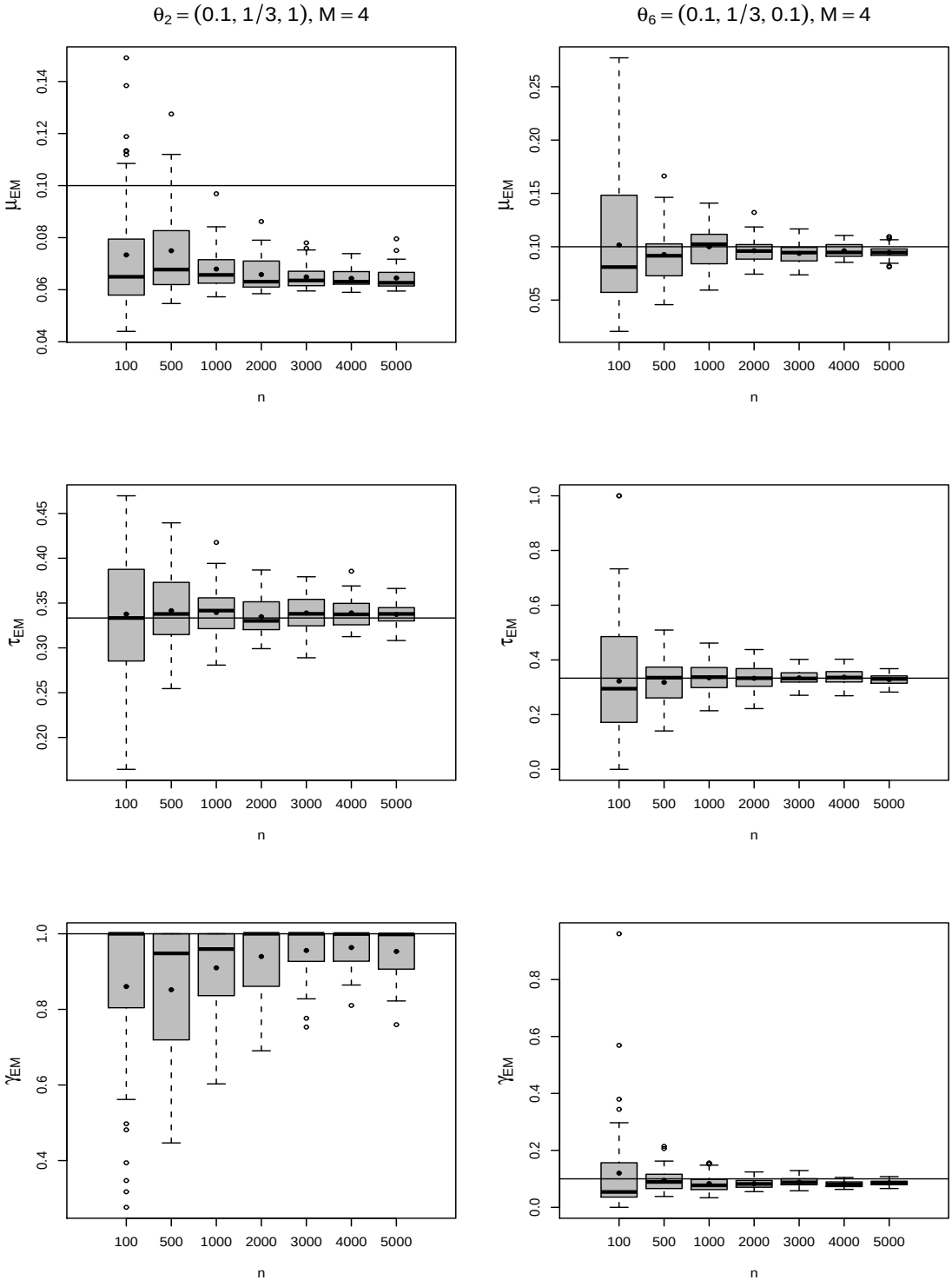


Figure 3.6 – Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_2 et θ_6 pour $M = 4$ et pour des valeurs croissantes de n.

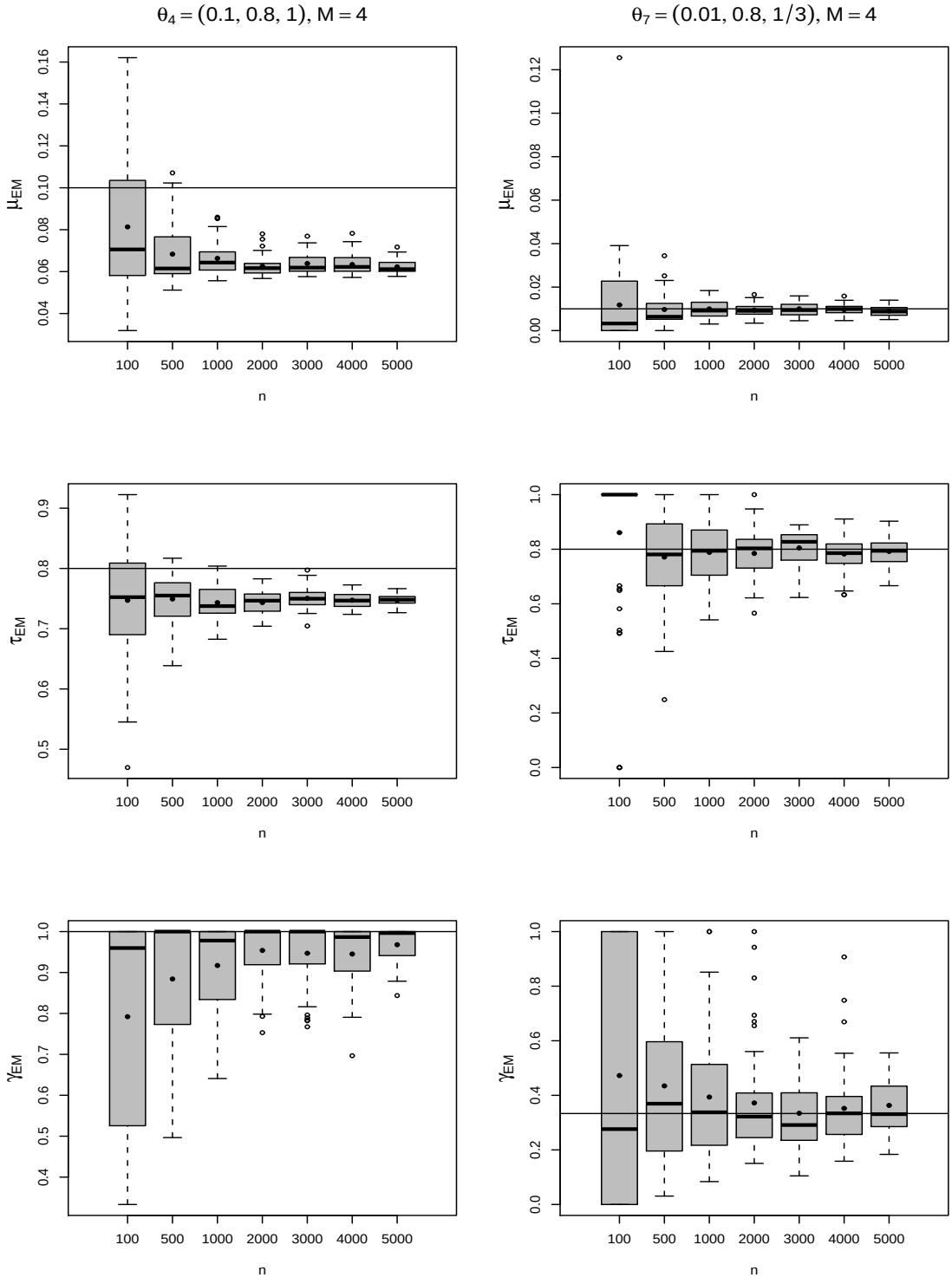


FIGURE 3.7 – Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_4 et θ_7 pour $M = 4$ et pour des valeurs croissantes de n.

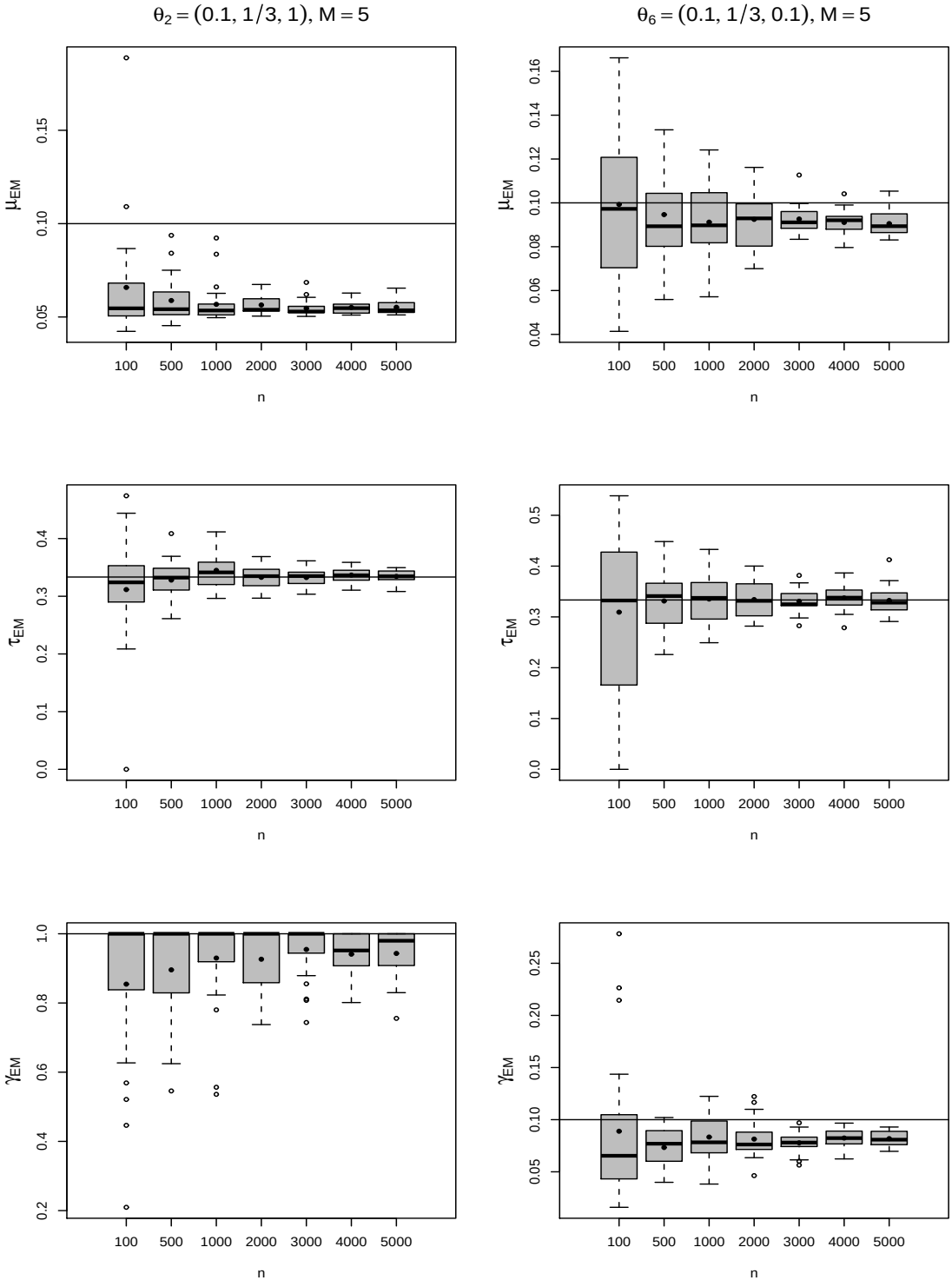


Figure 3.8 – Boîtes à moustaches de 25 valeurs estimées des trois coordonnées de θ_2 et θ_6 pour $M = 5$ et pour des valeurs croissantes de n.

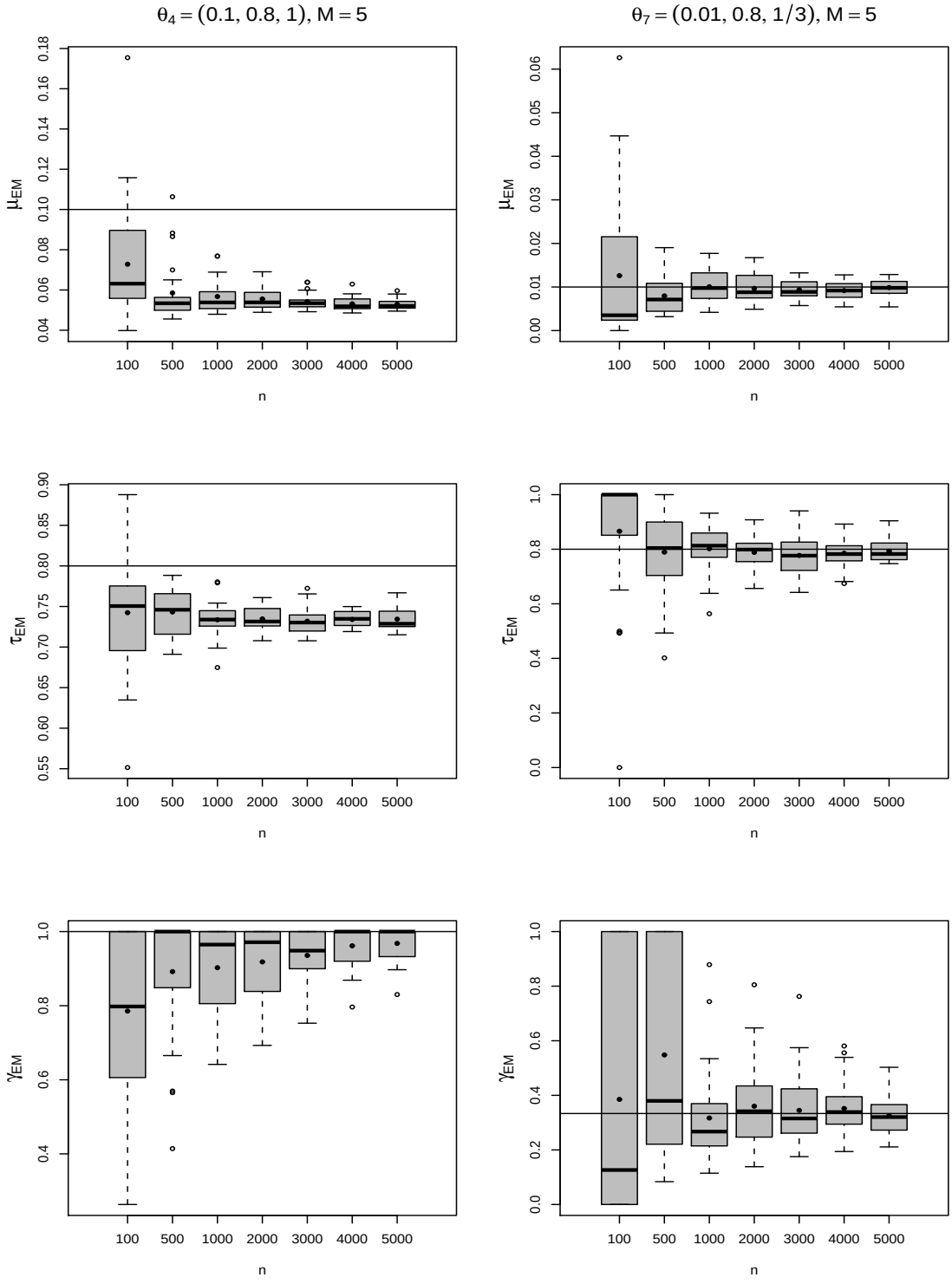


FIGURE 3.9 – Boîtes à moustaches de 25 valeurs estimées des trois coordonnées de θ_4 et θ_7 pour $M = 5$ et pour des valeurs croissantes de n.

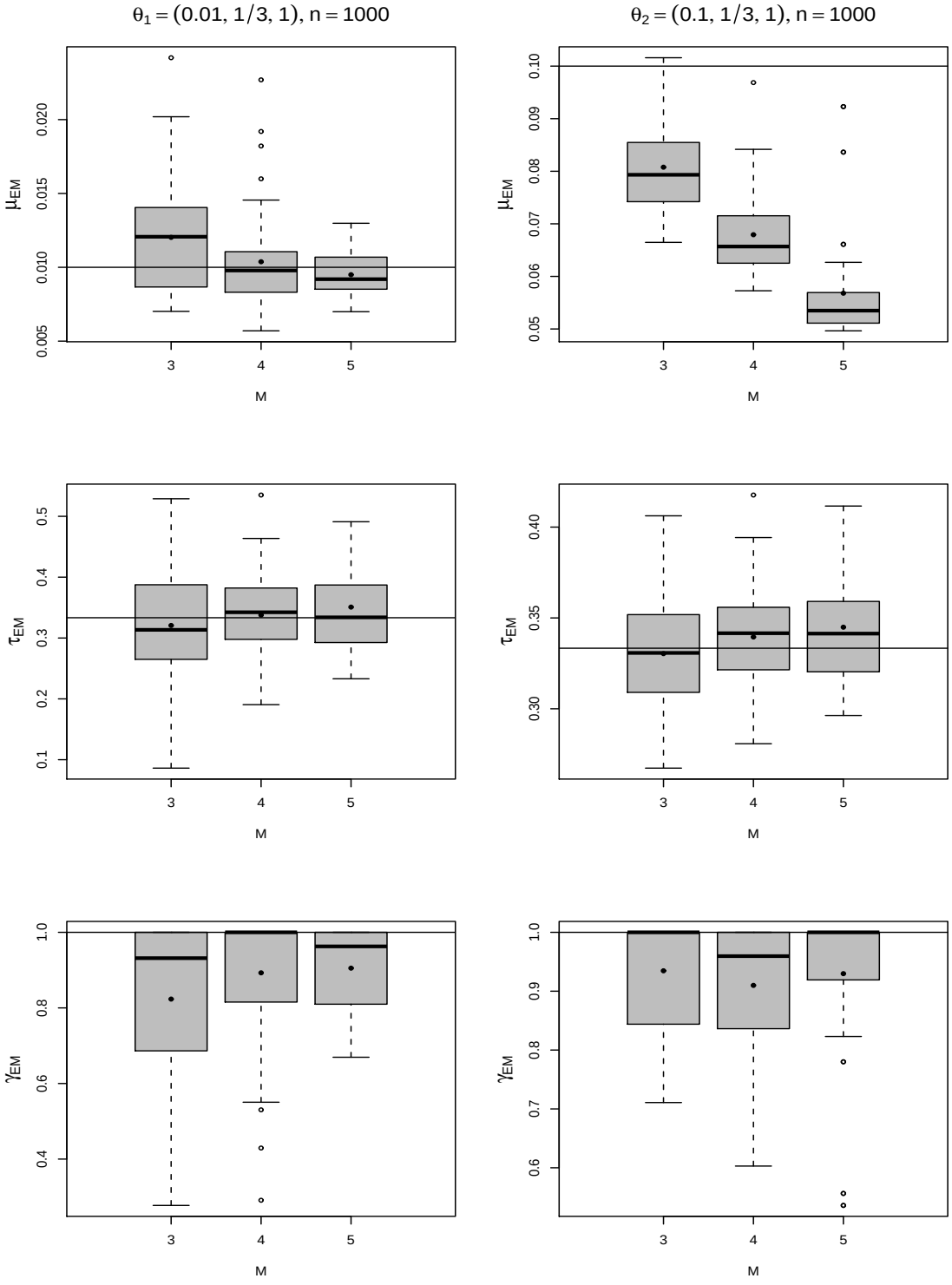


Figure 3.10 – Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_1 et θ_2 pour $n = 1\,000$ et $M \in \{3, 4\}$, et de 25 valeurs estimées pour $M = 5$.

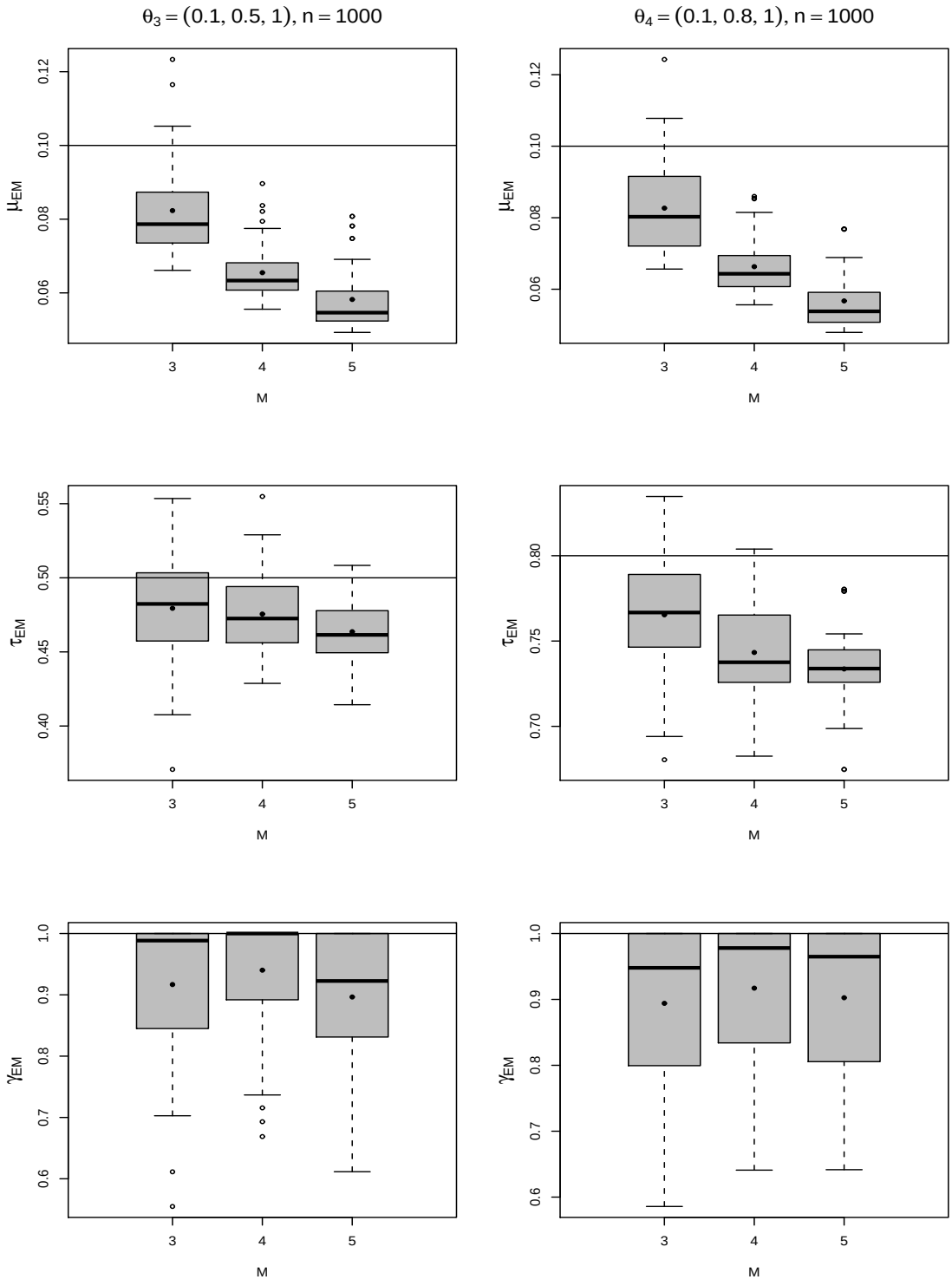


FIGURE 3.11 – Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_3 et θ_4 pour $n = 1\ 000$ et $M \in \{3, 4\}$, et de 25 valeurs estimées pour $M = 5$.

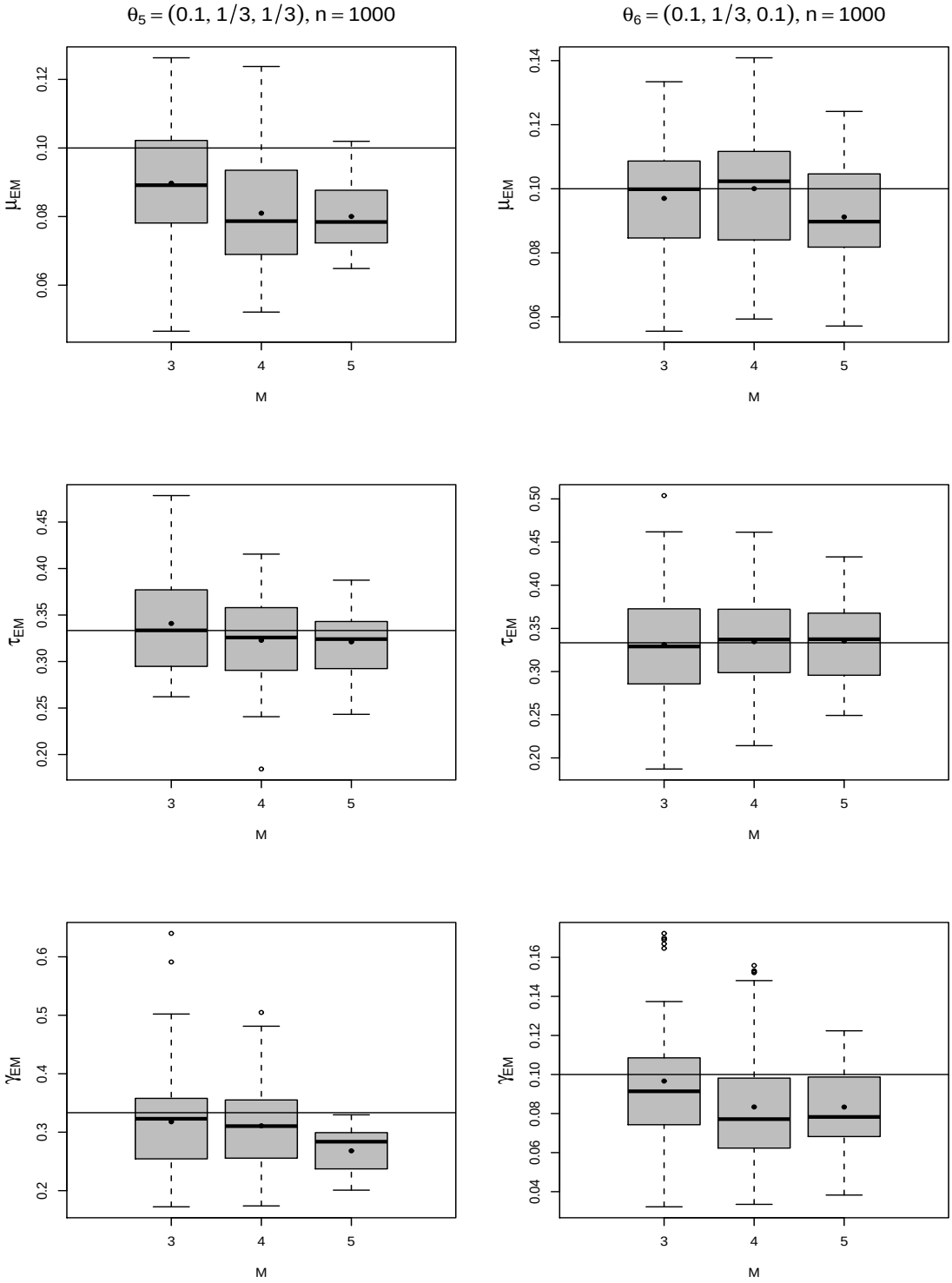


Figure 3.12 – Boîtes à moustaches de 50 valeurs estimées des trois coordonnées de θ_5 et θ_6 pour $n = 1\,000$ et $M \in \{3, 4\}$, et de 25 valeurs estimées pour $M = 5$.

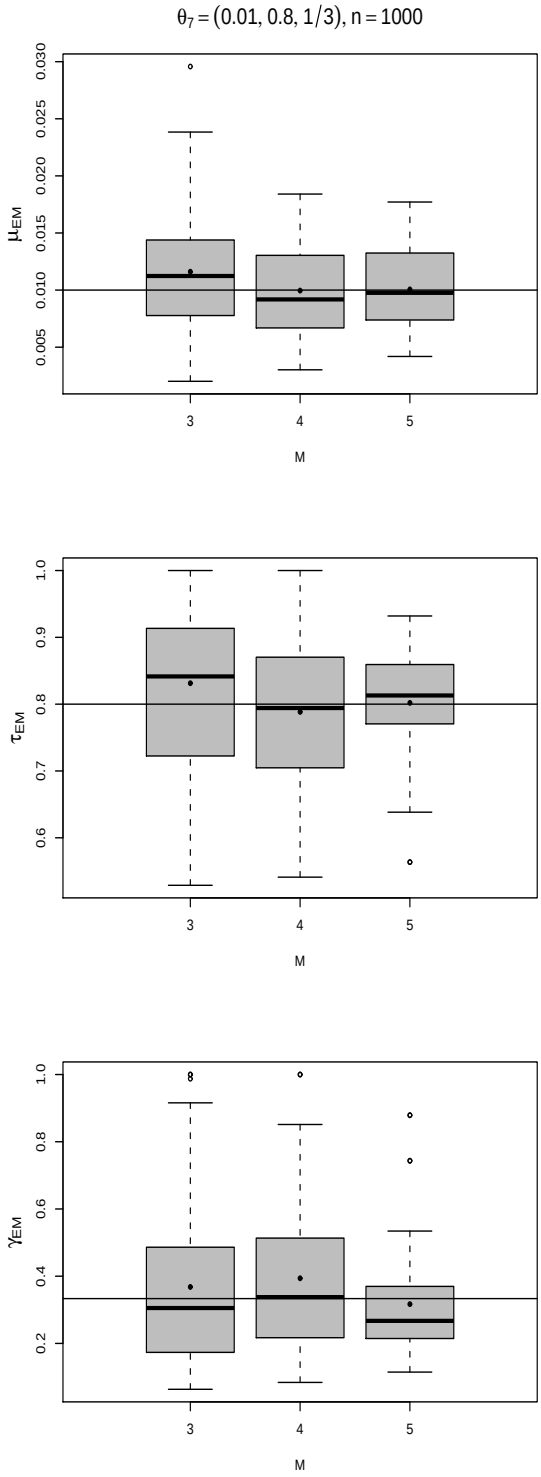


FIGURE 3.13 – Boîtes à moustaches de 50 valeurs estimées des trois coordonnées du paramètre θ_7 pour $n = 1\ 000$ et $M \in \{3, 4\}$, et de 25 valeurs estimées pour $M = 5$.

Résultats Les résultats obtenus sont résumés par deux types de graphiques. D'une part sont présentées dans les graphiques de 3.4 à 3.9 les boîtes à moustaches avec symbole de la moyenne de μ_{EM}, τ_{EM} et γ_{EM} pour les paramètres θ_2, θ_4, θ_6 et θ_7, avec M fixé et n variant de 100 à 5 000 pb. D'autre part figurent de 3.10 à 3.13 les boîtes à moustaches avec symbole de la moyenne de μ_{EM}, τ_{EM} et γ_{EM} pour chaque paramètre de la table 3.1, avec $n = 1\ 000$ pb et M variant de 3 à 5. Un trait horizontal représente pour chaque graphique la valeur μ_{V}, τ_{V} ou γ_{V}.

Les premiers graphiques montrent que de façon générale, la taille des boîtes diminue lorsque la longueur des séquences augmente traduisant la convergence de l'estimateur. Une longueur de séquences de 100 pb semble trop faible pour estimer correctement les paramètres. Les boîtes correspondant à τ_{EM} et γ_{EM} restent bien centrées autour de τ_{V} et γ_{V}, et nous en déduisons que la qualité de l'estimation est plutôt bonne. Nous remarquons que l'écart interquartile est en général plus faible pour τ_{EM} que pour γ_{EM}, et que les points aberrants sont également moins nombreux pour le deuxième paramètre que pour le troisième. L'estimation de τ_{V} semble ainsi meilleure que celle de γ_{V}. Nous observons en revanche un décalage de certaines boîtes de μ_{EM} qui se stabilisent en-dessous de μ_{V}. Ce phénomène traduit une sous-estimation de μ_{V}, pouvant être expliquée par une sous-estimation du nombre de substitutions au fil des étapes évolutives. En effet, même si les substitutions multiples sont autorisées, certaines substitutions inverses par exemple ne sont pas prises en compte. Ce phénomène est classique pour les modèles d'évolution.

La deuxième série de graphiques confirme la tendance de μ_{EM} à la sous-estimation de μ_{V}, celle-ci s'accentuant lorsque M augmente. Pour θ_1, θ_6 et θ_7 en revanche, nous n'observons pas de tel décalage. Nous observons de plus que la taille des boîtes, donc la dispersion, reste à peu près stable et même qu'elle diminue légèrement lorsque M croît.

Comparaison aux modèles de Kimura et de Jukes et Cantor Le taux global de substitution k_{EM} est défini à partir de μ_{EM} par l'équation suivante :

$$\mu_{\text{EM}} = \frac{2k_{\text{EM}}t}{M-1}. \tag{3.3.2}$$

En effet, les paramètres sont estimés à partir de deux séquences $\mathbf{X}(1)$ et $\mathbf{X}(M)$ ayant divergé depuis le temps t, et $2k_{\text{EM}}t$ est normalisé par le nombre d'étapes évolutives. Nous notons ns la proportion de sites qui diffèrent entre $\mathbf{X}(1)$ et $\mathbf{X}(M)$, ts la proportion de transitions séparant $\mathbf{X}(1)$ de $\mathbf{X}(M)$ et tv la proportion de transversions correspondante. Ainsi nous avons $ns = ts + tv$.

Pour confronter notre algorithme à des méthodes existantes dans le cas $\gamma_{\text{V}} = 1$, nous procédons à deux types de comparaisons. D'une part, nous comparons les deux premières coordonnées μ_{EM} et τ_{EM} des valeurs estimées par l'algorithme EM-HMM

avec les estimations données par la méthode de calcul direct de Kimura, figurant dans [50]. Soit α le taux de transition et β le taux de transversion définis dans le modèle de Kimura. La résolution d'un couple d'équations différentielles liées aux comptages des transitions et des transversions donne les valeurs suivantes des taux :

$$\begin{cases} \alpha = & \frac{1}{8T}\ln(1-2tv) - \frac{1}{4T}\ln(1-2ts-tv), \\ \beta = & -\frac{1}{8T}\ln(1-2tv). \end{cases}$$

Le taux de substitution global par site et par an estimé par le modèle de Kimura vaut $k_{\mathrm{KM}} = \alpha + 2\beta$. A l'aide de l'équation (3.3.2), nous pouvons comparer μ_{EM} et μ_{KM}, où :

$$\mu_{\mathrm{KM}} = \frac{2k_{\mathrm{KM}}t}{M-1} = -\frac{1}{2(M-1)}\ln((1-2ts-tv)\sqrt{1-2tv}).$$

De la même manière, nous pouvons comparer la proportion τ_{EM} de transitions estimée par EM-HMM à celle estimée par le modèle de Kimura, notée :

$$\tau_{\mathrm{KM}} = \frac{\alpha}{\alpha+2\beta} = \frac{-\ln(1-2ts-tv) + 1/2\ln(1-2tv)}{-\ln(1-2ts-tv) - 1/2\ln(1-2tv)}.$$

D'autre part, dans la situation neutre $(\tau_{\mathrm{V}}, \gamma_{\mathrm{V}}) = (1/3, 1)$, le modèle de référence est celui de Jukes et Cantor, introduit dans [47].

Définition 4. *La distance de Jukes-Cantor est définie par :*

$$K_{JC} = 2k_{JC}t = -\frac{3}{4}\ln\left(1-\frac{4}{3}ns\right). \tag{3.3.3}$$

Nous comparons μ_{EM} avec :

$$\mu_{\mathrm{JC}} = \frac{K_{\mathrm{JC}}}{M-1} = -\frac{3}{4(M-1)}\ln\left(1-\frac{4}{3}ns\right).$$

Dans la table 3.2 figurent les valeurs moyennes de μ_{EM}, μ_{KM}, μ_{JC}, τ_{EM} et τ_{KM}, estimées pour des répétitions de l'estimation du paramètre $\theta_2 = (0.1, 1/3, 1)$. Pour $M \in \{3, 4\}$, les répétitions sont au nombre de 100, et pour $M = 5$ seulement 25 répétitions ont été réalisées.

Les simulations nous montrent que les trois valeurs estimées du paramètre μ_{V} sont proches les unes des autres lorsque $M = 3$, mais qu'elles s'éloignent lorsque M augmente. Les estimations pour τ_{V} sont stables autour de la valeur $1/3$, même lorsque $M = 5$. Nous remarquons que les estimations de Kimura et de Jukes et Cantor sont ici toujours très proches l'une de l'autre.

M	rep	$\overline{\mu}_{\mathrm{EM}}$	$\overline{\mu}_{\mathrm{KM}}$	$\overline{\mu}_{\mathrm{JC}}$	$\overline{\tau}_{\mathrm{EM}}$	$\overline{\tau}_{\mathrm{KM}}$
3	100	0.082	0.107	0.107	0.349	0.331
4	100	0.067	0.109	0.109	0.368	0.337
5	25	0.058	0.109	0.109	0.327	0.325

Table 3.2 – Moyennes $\overline{\mu}_{\mathrm{EM}}$, $\overline{\mu}_{\mathrm{KM}}$, $\overline{\mu}_{\mathrm{JC}}$, $\overline{\tau}_{\mathrm{EM}}$ et $\overline{\tau}_{\mathrm{KM}}$ calculées sur rep répétitions des estimations μ_{EM}, μ_{KM}, μ_{JC}, τ_{EM} et τ_{KM}, pour θ_2 et $M \in \{3, 4, 5\}$.

3.3.4 Application à des données réelles

Mise en œuvre Lorsque nous considérons deux séquences homologues réelles $\mathbf{x}$ et $\mathbf{y}$, nous ignorons si l'une est l'ancêtre de l'autre. Ces deux séquences sont plus probablement deux séquences filles descendant d'une séquence racine inconnue $\mathbf{r}$. La vraisemblance s'écrit, en adaptant l'équation (1.1.4.3) :

$$L(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{r} \in \mathcal{A}^n} \Pi_{\mathbf{r}} \mathbb{P}_\theta(\mathbf{x}|\mathbf{r}) \mathbb{P}_\theta(\mathbf{y}|\mathbf{r}).$$

En pratique, le calcul de cette somme est numériquement coûteux, c'est pourquoi nous adoptons l'approche suivante. Si le modèle était réversible, la somme précédente se réduirait à $\Pi_{\mathbf{y}} \mathbb{P}_\theta(\mathbf{x}|\mathbf{y})$ ou $\Pi_{\mathbf{x}} \mathbb{P}_\theta(\mathbf{y}|\mathbf{x})$, selon que $\mathbf{y}$ est l'ancêtre de $\mathbf{x}$, ou $\mathbf{x}$ l'ancêtre de $\mathbf{y}$. Nous choisissons de rendre le modèle symétrique en posant :

$$L(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \left(\Pi_{\mathbf{x}} \mathbb{P}_\theta(\mathbf{y}|\mathbf{x}) + \Pi_{\mathbf{y}} \mathbb{P}_\theta(\mathbf{x}|\mathbf{y}) \right),$$

ce qui implique la réalisation de deux boucles Forward-Backward au lieu d'une seule, pour chaque étape E.

Comme l'algorithme ne donne des résultats dans un temps raisonnable que pour $M \in \{3, 4, 5\}$, nous limitons l'évolution à quatre intervalles de temps au maximum. Comme le modèle autorise plusieurs substitutions par intervalle de temps, cette limitation n'est pas pénalisante tant que le paramètre reste dans des limites raisonnables. Notons **Ma** pour un million d'années. Le taux de substitution moyen $\overline{k}$ dans les séquences biologiques a été estimé dans [63] pour diverses espèces. Il varie de 1% pour 50 Ma dans les régions à forte pression sélective, qui évoluent assez lentement, à 0.7% par Ma dans les régions à faible pression de sélection, qui évoluent plus rapidement.

Les simulations montrent que la limite inférieure pour l'estimation de μ est d'environ 10^{-2}, valeur au-delà de laquelle l'algorithme renvoie un résultat situé sur la frontière de Θ. La limite supérieure en temps de calcul apparaît pour une valeur de μ d'environ 0.2.

L'approximation de μ_{KM} par $2\overline{k}t/(M-1)$, où t est la durée depuis la divergence des deux séquences à partir d'un ancêtre commun, et $\overline{k}$ varie dans les limites données précédemment, nous fournit les valeurs de t minimales et maximales que nous pouvons considérer. Ainsi, t varie entre 10^8 et $2 \cdot 10^9$ années pour les régions évoluant lentement, et entre $3 \cdot 10^6$ et $6 \cdot 10^7$ pour les régions évoluant rapidement. Pour obtenir un intervalle de temps plus étendu, il faudrait pouvoir augmenter la valeur de M. En revanche, les séquences peuvent être longues, jusqu'à 10 000 pb. Plus elles le sont, meilleure est l'estimation.

Application à des séquences codantes Appliquons l'algorithme d'estimation à l'alignement des séquences nucléotidiques codantes de l'insuline chez la Souris et le Rat. La séquence codante d'un gène correspond à l'ensemble de ses exons mis bout à bout, et est appelée **CDS** pour Coding DNA Sequence. Les séquences sont disponibles dans les banques de données génomiques sous les identifiants respectifs $gi|52712 : 829-1155$ et $gi|56487 : 289-621$. Elles ont été recueillies sur GenBank, et ont été alignées à l'aide de ClustalW. Après la suppression des lacunes, l'alignement est de longueur $n = 327$ pb. Les résultats pour $M = 5$ sont les estimations suivantes :

$$\mu_{\mathrm{EM}} = 0.060, \quad \tau_{\mathrm{EM}} = 0.717, \quad \gamma_{\mathrm{EM}} = 0.241,$$

qui indiquent notamment que les transitions sont environ cinq fois plus fréquentes que les transversions et que l'évolution est fortement dépendante du contexte. Nous remarquons que l'estimation μ_{EM} est supérieure aux estimations obtenues par les modèles de Kimura et de Jukes et Cantor, alors que les estimations τ_{EM} et τ_{KM} sont très proches. En effet :

$$\mu_{\mathrm{JC}} = \mu_{\mathrm{KM}} = 0.023 \quad \text{et} \quad \tau_{\mathrm{KM}} = 0.762.$$

La durée depuis la divergence entre le Rat et la Souris est estimée dans [54] à environ 41 Ma. Nous pouvons ainsi estimer le taux global de substitution par :

$$k_{\mathrm{EM}} = \frac{\mu_{\mathrm{EM}}(M-1)}{2t} = \frac{0.06 \cdot 4}{2 \cdot 41 \cdot 10^6} = 3 \cdot 10^{-9}, \tag{3.3.4}$$

soit 0.3% par site et par Ma. Ce taux correspond à des régions à pression de sélection moyenne ou faible. La séquence de l'insuline est normalement assez conservée au fil du temps, et devrait plutôt avoir un taux de substitution correspondant à une pression sélective moyenne ou forte. Les taux k_{JC} et k_{KM} sont environ trois fois plus faibles que k_{EM}, mais restent élevés pour une telle région. Peut-être la séquence de 327 nucléotides est-elle un peu courte pour pouvoir estimer ce taux de façon fidèle. De plus, la valeur t est en dehors de l'intervalle recommandé pour les régions évoluant lentement.

Application à des séquences non codantes Appliquons maintenant cet algorithme à des alignements de séquences de pseudogènes d'hominidés. Les trois espèces considérées sont l'Homme, le Chimpanzé et le Gorille, et leur arbre phylogénétique est représenté à la figure 3.14. La distance de Jukes-Cantor (3.3.3) est définie au paragraphe 3.3.3.

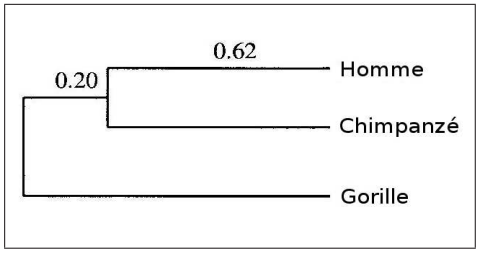


Figure 3.14 – Arbre phylogénétique reliant l'Homme, le Chimpanzé et le Gorille. Les longueurs des branches sont données par les distances de Jukes-Cantor.

Le pseudogène étudié est la séquence de la bêta-globine, une protéine présente chez tous les grands singes. Les trois séquences sont disponibles dans les banques de données génomiques sous les identifiants respectifs $gi|29444$, $gi|38204$ et $gi|22875$. Elles ont été recueillies sur GenBank, et ont été alignées deux à deux à l'aide de ClustalW. Après la suppression des lacunes, les trois alignements Homme-Chimpanzé, Chimpanzé-Gorille et Gorille-Homme sont de longueur respective $n = 2\ 147$, $2\ 145$ et $2\ 147$ pb. Les résultats pour $M = 5$ sont pour l'alignement Homme-Chimpanzé :

$$\mu_{\mathrm{EM}} = 0.004, \quad \tau_{\mathrm{EM}} = 0.634, \quad \gamma_{\mathrm{EM}} = 0.839,$$

pour l'alignement Chimpanzé-Gorille :

$$\mu_{\mathrm{EM}} = 0.015, \quad \tau_{\mathrm{EM}} = 0.768, \quad \gamma_{\mathrm{EM}} = 0.399,$$

et pour l'alignement Gorille-Homme :

$$\mu_{\mathrm{EM}} = 0.007, \quad \tau_{\mathrm{EM}} = 0.702, \quad \gamma_{\mathrm{EM}} = 0.507.$$

Ces valeurs indiquent que les transitions sont trois à sept fois plus fréquentes que les transversions selon l'alignement considéré, et que la dépendance au contexte est moyenne. Les modèles de Kimura et de Jukes et Cantor donnent respectivement les valeurs suivantes :

$$\begin{aligned} \mu_{\mathrm{JC}} = \mu_{\mathrm{KM}} = 0.004 \qquad &\text{et} \qquad \tau_{\mathrm{KM}} = 0.647, \\ \mu_{\mathrm{JC}} = \mu_{\mathrm{KM}} = 0.005 \qquad &\text{et} \qquad \tau_{\mathrm{KM}} = 0.775, \\ \mu_{\mathrm{JC}} = \mu_{\mathrm{KM}} = 0.004 \qquad &\text{et} \qquad \tau_{\mathrm{KM}} = 0.717. \end{aligned}$$

Nous remarquons que les deux estimations du paramètre τ sont toujours très proches. Les estimations du premier paramètre sont elles aussi assez proches, sauf pour l'alignement Chimpanzé-Gorille.

La durée depuis la divergence entre l'Homme est le Chimpanzé est estimée sur un ensemble de séquences non codantes dans [18], et vaut environ 4.7 Ma. Le Gorille a divergé du groupe Homme-Chimpanzé depuis environ 7.7 Ma. Nous pouvons ainsi estimer le taux global de substitution par la formule (3.3.4) à $k_{\mathrm{EM}} = 0.17\%$ par site et par Ma pour l'alignement Homme-Chimpanzé, 0.4% par site et par Ma pour l'alignement Chimpanzé-Gorille et 0.18% par site et par Ma pour l'alignement Gorille-Homme. Les pseudogènes sont peu soumis à la sélection naturelle, et évoluent ainsi de façon plus rapide que les séquences codantes qui sont mieux conservées. Nous retrouvons ce phénomène ici, puisque les trois taux de substitution sont de valeur moyenne, avec un taux plutôt élevé à 0.4% pour l'alignement Chimpanzé-Gorille. Les taux k_{JC} et k_{KM} sont en moyenne plus faibles que k_{EM}.

3.4 Perspectives

Le modèle de dépendance à gauche que nous venons de décrire peut se généraliser de différentes manières. Nous pouvons tout d'abord étendre le modèle au cas de trois séquences homologues et plus. D'autre part, pour étendre la portée de la dépendance, nous pouvons considérer des codons à la place des nucléotides, ce qui porte l'alphabet à 61 codons non-stop au lieu de 4 nucléotides. Cette approche engendre de ce fait des problèmes numériques liés au temps de calcul et à la mémoire nécessaires à l'algorithme d'estimation. Enfin, pour revenir dans le cadre du modèle de triplets, la dépendance peut être élargie au voisin de droite, ce qui nous amène à considérer des champs markoviens à la place des processus markoviens.

Deuxième partie

Tests d'hypothèses pour tables de contingence creuses

Chapitre 4

Motivation biologique et notations

L'observation de phénomènes physiques, sociologiques ou biologiques, comme ceux exposés dans le chapitre 1 conduit souvent au recueil de données sous forme de tableaux à entrées multiples correspondant aux différents caractères d'intérêt. Les tables de dénombrement correspondantes sont appelées **tables de contingence**. Ainsi, les individus sont classés pour chaque caractère dans l'une des catégories possibles choisies préalablement par l'observateur, et chaque case du tableau contient le nombre d'individus appartenant à l'intersection des catégories formant ses coordonnées. L'analyse de l'information consiste à se demander si certains caractères ou certaines combinaisons de caractères sont indépendants. Le test le plus répandu pour répondre à ces questions est le test du khi-deux d'indépendance, introduit par Pearson dans [64]. Ce test découle d'un résultat asymptotique, le théorème de la limite centrale.

En pratique pour que le test soit valide et que nous puissions accepter le résultat théorique, les effectifs théoriques sous l'hypothèse d'indépendance doivent être « assez grands » : nous les supposons en général plus grands que 5. Le test n'est donc pas utilisable lorsque la table comporte des cases à effectifs nuls nommés **zéros**, car il y a alors de grandes chances que les effectifs théoriques soient très faibles. Une table comportant des zéros est appelée table **creuse**. La technique du regroupement des classes n'est pas toujours satisfaisante et elle n'a pas de sens dans le cas des données que nous considérons.

Des corrections existent pour surmonter ce problème de tables creuses, mais elles ne sont pas efficaces lorsque les cases nulles sont nombreuses. Le test exact de Fisher par exemple, valable pour de petits effectifs théoriques, devient très vite numériquement inutilisable lorsque l'effectif total, le nombre de caractères ou le nombre de catégories augmente. D'autres statistiques de même loi asymptotique que celle de la **statistique de Pearson** ont été introduites, par exemple la **statistique de Kullback** du Minimum d'Information Discriminante dans [40] ainsi que la famille générale de statistiques de Read et Cressie dans [26]. Une vue d'ensemble de ces

travaux est donnée dans [49, 51]. Notre objectif est de mettre en place, en corrigeant les statistiques existantes, des tests d'indépendance de caractères dans de grandes tables de contingence multidimensionnelles possédant de nombreux zéros.

Pour cela, nous choisissons de travailler de façon plus générale sur les vecteurs d'effectifs et les tests d'adéquation à une loi multinomiale. Nous commençons par présenter le problème biologique qui a motivé notre travail. Nous relions ensuite vecteurs d'effectifs et tables de contingence, et nous introduisons les principales notations qui en découlent. Enfin, nous exposons les problèmes que posent les tables ou vecteurs creux. Ce chapitre, ainsi que les chapitres 5 et 6, représentent la version détaillée de l'article [32].

4.1 Motivation biologique

L'exemple principal qui nous sert de motivation pour la deuxième partie de ce travail, est celui du test de l'indépendance mutuelle des trois caractères d'une table de contingence obtenue à l'aide de données génomiques. Les codons sont les unités élémentaires des séquences codantes, et la dépendance dans les évolutions des trois positions est claire. Nous considérons par conséquent des triplets de positions dans le matériel génétique non codant. Nous construisons ensuite un test d'hypothèses permettant de décider si l'évolution y est dépendante ou non.

Ce test est appliqué dans le chapitre 6 à des tables de contingence en dimension trois qui ont la particularité de présenter un grand nombre de cases d'effectif nul, que nous décrivons ci-dessous. La théorie du test que nous présentons est générale et ne dépend pas de la structure des données considérées ni de la dimension de la table de contingence. C'est la raison pour laquelle le test peut également être utilisé dans des cadres variés dont nous donnerons quelques exemples au chapitre 6.

Nous considérons ainsi de façon générale un alignement sans lacunes de deux séquences $\mathbf{y}$ et $\mathbf{z}$ de même longueur $3n$, où les sites sont regroupés par triplets non-chevauchants. Ces n triplets peuvent représenter les codons d'un exon, mais aussi des sites quelconques d'une séquence intergénique. Nous n'autorisons pas les chevauchements de triplets voisins parce que nous procédons à des tests d'hypothèses à partir de ces triplets, et les observations considérées doivent être indépendantes et identiquement distribuées. Notons les séquences :

$$\mathbf{y} = (y_1, y_2, \ldots, y_n) \text{ et } \mathbf{z} = (z_1, z_2, \ldots, z_n),$$

avec $y_i = (y_i^1, y_i^2, y_i^3)$ et $z_i = (z_i^1, z_i^2, z_i^3)$ pour tout i dans $\{1, \ldots, n\}$. Elles sont alignées comme suit :

$$\begin{array}{ccccccccccccc} y_1^1 & y_1^2 & y_1^3 & , & y_2^1 & y_2^2 & y_2^3 & , & \cdots & , & y_n^1 & y_n^2 & y_n^3 \\ z_1^1 & z_1^2 & z_1^3 & , & z_2^1 & z_2^2 & z_2^3 & , & \cdots & , & z_n^1 & z_n^2 & z_n^3 \end{array}.$$

L'hypothèse biologique que nous voulons tester pour justifier le développement de modèles d'évolution avec dépendance au contexte est l'indépendance de l'évolution des trois positions d'un triplet, donc l'indépendance au sein des colonnes $\binom{y_i}{z_i}$. C'est pourquoi nous nous plaçons dans le cadre des tables de contingence. La table associée aux données est de dimension trois, parce que chaque position du triplet est un caractère d'intérêt. Les catégories sont au nombre de $R = IJK$ avec $I = J = K = 16$. Pour chacune des trois dimensions, les $4^2 = 16$ colonnes distinctes ou motifs possibles sont numérotées par ordre lexicographique de 1 à 16 :

$$\binom{A}{A}, \binom{A}{C}, \binom{A}{G}, \binom{A}{T}, \binom{C}{A}, \binom{C}{C}, \binom{C}{G}, \binom{C}{T}, \binom{G}{A}, \binom{G}{C}, \binom{G}{G}, \binom{G}{T}, \binom{T}{A}, \binom{T}{C}, \binom{T}{G}, \binom{T}{T}.$$

Nous complétons la table en comptant dans la suite des colonnes :

$$\left\{ \binom{y_i}{z_i}, i \in \{1, \ldots, n\} \right\}$$

le nombre d'occurrences de chacun des $16^3 = 4\ 096$ triplets de colonnes possibles :

$$\binom{y_i^1}{z_i^1}\binom{y_i^2}{z_i^2}\binom{y_i^3}{z_i^3}.$$

Ainsi, la case de coordonnées (i, j, k) dans la table contient le nombre n_{ijk} d'apparitions le long de la séquence du motif (i, j, k). Par exemple en position $(7, 2, 15)$, nous trouvons le nombre d'occurrences de :

$$\binom{C}{G}\binom{A}{C}\binom{T}{G}.$$

4.2 Notations pour les tables de contingence

En raison de notre motivation biologique, les résultats sont donnés pour des tables de contingence de dimension trois. Les notations sont aisément généralisables à d'autres dimensions. Nous notons $\mathcal{I}$, $\mathcal{J}$ et $\mathcal{K}$ les caractères étudiés. Nous supposons que les unités de l'échantillon de taille n étudié sont décrites par leur appartenance à l'une des I catégories (ou modalités) du caractère $\mathcal{I}$, l'une des J catégories du caractère $\mathcal{J}$ et l'une des K catégories du caractère $\mathcal{K}$. Soit $R = IJK$ le nombre de cases de la table de contingence : il représente le nombre total de classes possibles. Nous disposons avec cet n-échantillon d'une réalisation du vecteur aléatoire de taille R répartissant les n individus dans les R classes selon une loi multinomiale dont les paramètres sont inconnus.

Voici deux représentations possibles des données qui ont chacune leurs avantages. Les effectifs observés sont alors notés n_{ijk} pour $1 \leqslant i \leqslant I$, $1 \leqslant j \leqslant J$ et $1 \leqslant k \leqslant K$,

ou n_r pour $1 \leqslant r \leqslant R$ selon la nature de la représentation, et sont des réalisations des variables aléatoires N_{ijk} ou N_r.

La **première représentation** possible est une table de contingence en dimension trois, s'écrivant par exemple sous la forme de K tables en dimension deux indexées par k dans $\{1, \dots, K\}$:

I \ **J**	**1**	$\cdots$	**j**	$\cdots$	**J**	**Total**
1	n_{11k}	$\cdots$	n_{1jk}	$\cdots$	n_{1Jk}	n_{1+k}
2	n_{21k}	$\cdots$	n_{2jk}	$\cdots$	n_{2Jk}	n_{2+k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1k}	$\cdots$	n_{ijk}	$\cdots$	n_{iJk}	n_{i+k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1k}	$\cdots$	n_{Ijk}	$\cdots$	n_{IJk}	n_{I+k}
Total	n_{+1k}	$\cdots$	n_{+jk}	$\cdots$	n_{+Jk}	n_{++k}

En sommant les effectifs sur les lignes, colonnes et hauteurs, nous définissons les effectifs marginaux à deux indices qui apparaissent dans les marges de la table :

$$\forall (i,k) \in \{1,\dots,I\} \times \{1,\dots,K\}, \qquad n_{i+k} = \sum_{j=1}^{J} n_{ijk},$$

$$\forall (i,j) \in \{1,\dots,I\} \times \{1,\dots,J\}, \qquad n_{ij+} = \sum_{k=1}^{K} n_{ijk},$$

$$\forall (j,k) \in \{1,\dots,J\} \times \{1,\dots,K\}, \qquad n_{+jk} = \sum_{i=1}^{I} n_{ijk},$$

puis en sommant ces marges, nous obtenons les effectifs marginaux à un indice :

$$\forall i \in \{1,\dots,I\}, \qquad n_{i++} = \sum_{j=1}^{J} n_{ij+} = \sum_{k=1}^{K} n_{i+k} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk},$$

$$\forall j \in \{1,\dots,J\}, \qquad n_{+j+} = \sum_{k=1}^{K} n_{+jk} = \sum_{i=1}^{I} n_{ij+} = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{ijk},$$

$$\forall k \in \{1,\dots,K\}, \qquad n_{++k} = \sum_{i=1}^{I} n_{i+k} = \sum_{j=1}^{J} n_{+jk} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ijk}.$$

Enfin, ces marges doubles sont telles que :

$$\sum_{i=1}^{I} n_{i++} = \sum_{j=1}^{J} n_{+j+} = \sum_{k=1}^{K} n_{++k} = n.$$

La **seconde représentation** est un vecteur x de taille R obtenu en alignant toutes les cases de la table selon une numérotation choisie pour les R cases :

$$x = (n_1, \ldots, n_R).$$

Les tables se prêtent bien aux tests d'indépendance des caractères, parce qu'elles permettent de visualiser directement les hypothèses sur les lignes, colonnes et hauteurs. Il est possible de tester l'indépendance des caractères, qu'elle soit mutuelle, partielle ou conditionnelle. Les tables deviennent en revanche plus difficiles à manipuler en dimension supérieure. Il existe des présentations astucieuses de telles tables, mais elles reviennent en général à introduire une asymétrie dans les caractères en en privilégiant certains. Cette approche est valable pour des données issues de questionnaires de sondage par exemple, mais elle semble artificielle lorsque nous considérons trois positions arbitraires d'une séquence biologique.

Les vecteurs sont quant à eux les supports standards pour les résultats théoriques. Les tests d'indépendance de caractères dans les tables de contingence sont en effet des cas particuliers des tests d'adéquation aux lois multinomiales pour lesquelles nous présentons quelques rappels. Les modèles d'échantillonnage utilisés pour remplir les tables sont les suivants.

- Sans aucune contrainte d'effectif, le **modèle** est **de Poisson**. Ce modèle intervient par exemple lorsque l'expérimentateur décide d'observer les occurrences d'un phénomène pendant une durée fixée. Le nombre d'observations est donc une variable aléatoire.
- Lorsque seul le nombre n d'observations est fixé, le **modèle** considéré est **multinomial** et consiste à répartir les n individus dans les R catégories. Lors de certaines enquêtes, le nombre de personnes à interroger est fixé par le concepteur de l'étude, mais il ne se préoccupe pas de leur représentativité par rapport à la population générale.
- Si certaines marginales sont fixées, le modèle est appelé **modèle multinomial indépendant**. Parfois, le sondage s'effectue par strates selon un critère choisi, et le nombre de personnes à interroger dans chaque strate est fixé.
- Enfin, si toutes les marginales sont imposées, le **modèle** est dit **hypergéométrique multivarié**.

De par la nature de nos applications, nous nous restreignons aux modèles multinomial et multinomial indépendant.

4.3 Modèle multinomial

Soit $Z = (Z_1, \ldots, Z_R)$ un vecteur aléatoire de taille R dont les réalisations prennent leurs valeurs $z = (z_1, \ldots, z_R)$ dans $\{0,1\}^R$ avec $\sum_{r=1}^R z_r = 1$. Un tel vecteur est appelé indicatrice multinomiale ou indicatrice de Bernoulli multivariée. Soit $p = (p_1, \ldots, p_R)$ une loi de probabilité quelconque sur $[0,1]^R$. Si nous définissons pour tout z dans $\{0,1\}^R$:

$$\mathbb{P}(Z = z) = \mathbb{P}(Z_1 = z_1, \ldots, Z_R = z_R) = p_1^{z_1} \ldots p_R^{z_R},$$

nous avons :

$$\sum_{z \in \{0,1\}^R} \mathbb{P}(Z = z) = \sum_{r=1}^R p_r = 1.$$

Soit $z^{1,n} = (z^1, \ldots, z^n)$ une réalisation d'un n-échantillon indépendant de même loi que Z. La vraisemblance de cet échantillon vaut :

$$L_{z^{1,n}}(p) = \prod_{j=1}^n \mathbb{P}(Z^j = z^j) = \prod_{j=1}^n p_1^{z_1^j} \ldots p_R^{z_R^j} = p_1^{\sum_{j=1}^n z_1^j} \ldots p_R^{\sum_{j=1}^n z_R^j}.$$

Notons $n_r = \sum_{j=1}^n z_r^j$ le nombre d'éléments de $z^{1,n}$ pour lesquels $z_r = 1$. Nous pouvons alors écrire :

$$L_{z^{1,n}}(p) = p_1^{n_1} \ldots p_R^{n_R} = \exp\left(\sum_{r=1}^R n_r \ln p_r\right).$$

Soit enfin $X = (N_1, \ldots, N_R)$ le vecteur aléatoire de réalisation $x = (n_1, \ldots, n_R)$. Ce vecteur suit la loi multinomiale de paramètres n et p :

Définition 5. *Soit n dans $\mathbb{N}^*$. Soit $p = (p_1, \ldots, p_R)$ un vecteur de $[0,1]^R$ tel que $p_1 + \cdots + p_R = 1$. Le vecteur aléatoire $X = (N_1, \ldots, N_R)$ à valeurs dans $\mathbb{N}^R$ suit une* **loi multinomiale** *de paramètres n et p notée $\mathcal{M}(n; p)$ lorsque pour tout vecteur $x = (n_1, \ldots, n_R)$ de $\{0, \ldots, n\}^R$ tel que $n_1 + \cdots + n_R = n$, nous avons :*

$$\mathbb{P}(X = x) = \mathbb{P}(N_1 = n_1, \ldots, N_R = n_R) = \frac{n!}{n_1! \ldots n_R!} p_1^{n_1} \ldots p_R^{n_R}.$$

Propriété 3. *Si le vecteur X suit la loi $\mathcal{M}(n; p)$, toutes les lois marginales et conditionnelles associées à X sont aussi des lois multinomiales. En particulier, la loi marginale de chacune de ses composantes N_r est une loi binomiale $\mathcal{B}(n; p_r)$.*

La vraisemblance d'un vecteur multinomial de dimension R d'effectif total n, ou de façon équivalente d'une table de contingence à R cases et à n observations s'écrit donc :

$$L_x(p) = p_1^{n_1} \dots p_R^{n_R},$$

et la log-vraisemblance correspondante :

$$\ell_x(p) = \ln(L_x(p)) = \sum_{r=1}^{R} n_r \ln(p_r).$$

Nous rappelons, sans démonstration, la propriété suivante.

Propriété 4. *L'estimateur du maximum de vraisemblance p^* du paramètre p pour des effectifs observés $x = (n_1, \dots, n_R)$ est convergent et vaut :*

$$p^* = \left(\frac{n_1}{n}, \dots, \frac{n_R}{n}\right).$$

La convergence de cet estimateur provient de la distribution binomiale des composantes N_r, et de l'application de l'inégalité de Bienaymé-Chebychev.

4.4 Tables et vecteurs creux

Une case d'effectif nul dans un échantillon multinomial pour ou dans une table de contingence issue du modèle multinomial correspondant, est appelée zéro. Pour une position r dans $\{1, \dots, R\}$ du vecteur d'effectifs x, le zéro peut être de deux types :

- **structural** : $n_r = 0$ et $p_r = 0$,
- **aléatoire** : $n_r = 0$ mais $p_r > 0$.

Les zéros structuraux sont imposés par la nullité de la probabilité d'appartenance à la catégorie r, alors que les zéros aléatoires apparaissent au hasard des tirages. Un zéro aléatoire en case r apparaît en général lorsque la probabilité p_r est petite et que l'échantillon n'est pas assez grand pour qu'au moins une observation appartienne à la classe r. Si l'échantillon était assez grand, une observation au moins appartiendrait à la catégorie r, ce qui n'est pas le cas pour des zéros structuraux. Pour distinguer zéros aléatoires et zéros structuraux, il faut analyser les données et vérifier pour chaque case s'il est possible d'appartenir à la catégorie correspondante. Si c'est le cas, le zéro est aléatoire, sinon il est structural.

Pour traiter les zéros structuraux, il suffit d'enlever les cases correspondant à ces zéros au vecteur de comptage, comme suggéré dans [2, 21]. Un vecteur sans zéros structuraux est dit **complet**. Ainsi, le seul type de zéros que nous considérons dans ce travail est celui des zéros aléatoires, c'est pourquoi nous supposerons toujours que

le vecteur est complet et donc que pour tout r dans $\{1, \ldots, R\}$, la probabilité p_r est strictement positive.

Lorsque la distribution multinomiale n'est pas trop asymétrique en faveur de certaines catégories, et lorsque l'échantillon considéré est grand, les zéros aléatoires sont peu nombreux. En revanche, dans le cas de lois très asymétriques, il est fréquent d'avoir de nombreux zéros aléatoires malgré un échantillon de taille importante et malgré des effectifs théoriques non nuls, mais proches de 0. Lorsqu'un vecteur ou une table possèdent de nombreux zéros, ils sont qualifiés de vecteur creux et table creuse.

Nous détaillerons dans les applications l'exemple présenté dans la section 4.1, où les tables creuses sont fréquentes malgré un nombre d'observations très grand. Ce phénomène est dû au grand nombre de catégories par rapport au nombre d'observations. La quantité n/r peut être définie comme un indicateur qui mesure à quel point la table est creuse, parce qu'elle représente le nombre moyen d'observations par case dans le cas équidistribué. Nous verrons pourquoi les tables creuses posent des problèmes dans l'application des tests d'hypothèses classiques, et en quoi nous pouvons y remédier par la construction de nouvelles statistiques de test.

Dans le chapitre 5, nous présentons les statistiques de test classiques pour tester l'adéquation à une distribution multinomiale. À partir de maintenant, nous adoptons la notation sous forme de vecteurs pour exposer les résultats théoriques, et ce jusqu'au développement des applications dans le chapitre 6.

Chapitre 5

Statistiques d'adéquation

Nous rappelons au début de ce chapitre les principales statistiques utilisées dans les tests d'adéquation, ainsi que leur domaine de validité en présence de faibles effectifs théoriques. Dans le cas des tables creuses, nous présentons une correction proposée par Ku pour la statistique de Kullback et ses limitations. Puis nous construisons une statistique de Kullback corrigée inspirée par le raisonnement de Ku, appropriée à une table possédant plusieurs zéros. Cette correction étant très faible, nous proposons de nouvelles corrections, à la fois pour la statistique de Pearson et pour la statistique de Kullback. Nous montrons que les statistiques corrigées et non corrigées ont même loi asymptotique, et étudions leur comportement en fonction du nombre de zéros. Nous observons que ces statistiques corrigées sont bien adaptées lorsque les zéros sont nombreux, c'est-à-dire lorsqu'ils représentent au moins la moitié des effectifs des catégories. Ce fait est confirmé par le calcul de risques de première espèce et de puissances empiriques.

5.1 Statistiques de tests d'adéquation à des lois multinomiales

Nous nous plaçons à présent dans le cadre d'un vecteur $x = (n_1, \ldots, n_R)$ d'effectif total n, réalisation d'un vecteur aléatoire X de loi multinomiale $\mathcal{M}(n; p)$ de paramètre inconnu p. Nous testons l'adéquation de x à une loi multinomiale $\mathcal{M}(n; p^0)$ pour p^0 un vecteur de probabilités théoriques sous $\mathcal{H}_0$:

$$\mathcal{H}_0 : \quad \mathcal{L}(X) = \mathcal{M}(n; p^0) \quad \text{contre} \quad \mathcal{H}_1 : \quad \mathcal{L}(X) = \mathcal{M}(n; q) \text{ avec } q \neq p^0.$$

Nous supposons que p^0 est une fonction connue dépendant d'un paramètre multidimensionnel θ appartenant à un ensemble Θ de $\mathbb{R}^s$, θ étant inconnu de l'observateur :

$$p^0 = p^0(\theta) \quad \text{avec} \quad \theta = (\theta_1, \ldots, \theta_s).$$

Si $s = 0$, p^0 est supposé connu et il n'y a alors aucun paramètre supplémentaire à estimer. Nous supposons que s est strictement inférieur à $R - 1$ avec $R \geqslant 2$.

Lorsque s est non nul, nous sommes amenés à écrire les statistiques de test à l'aide d'une estimation $\widehat{p^0(\theta)}$ de p^0, obtenue grâce à l'estimation $\hat{\theta}$ de θ. Ici nous supposons que $\widehat{p^0(\theta)} = p^0(\hat{\theta})$ car pour les tests que nous considérons, l'application $\theta \mapsto p^0(\theta)$ est bijective. Le plus souvent, $p^0(\theta)$ est estimé par la méthode du maximum de vraisemblance et nous notons la valeur obtenue p^{*0}. Nous omettrons ainsi toujours les indices n par souci de simplicité, même si tous les estimateurs que nous considérons ici dépendent de n.

Comme p est inconnu en dehors du cas de simulations exécutées par l'utilisateur, les probabilités p_r sont estimées par leur estimateur du maximum de vraisemblance $p_r^* = n_r/n$. Il y a donc deux estimateurs du maximum de vraisemblance : l'un, p^{*0}, sous l'hypothèse $\mathcal{H}_0$, et l'autre, p^*, pour le paramètre réel qui a engendré les données. Nous verrons au chapitre 6 comment traiter les données de façon à ce qu'aucun p_r^{*0}, $r \in \{1, \ldots, R\}$ ne soit nul, ce que nous supposons donc dans tout ce chapitre. Nous verrons également que nous pouvons faire d'autres choix d'estimateurs pour p et $p^0(\theta)$.

Plusieurs statistiques de test ont été développées au vingtième siècle. Elles se fondent sur des aspects différents des modèles mis en œuvre, et permettent de varier l'approche statistique selon la nature et la taille des données étudiées. Nous verrons qu'elles possèdent l'avantage d'avoir approximativement le même comportement asymptotique, ce qui nous permettra de comparer les tests entre eux pour les problèmes qui nous intéressent. Nous commençons par présenter une famille générale, la famille de statistiques de divergence de puissance, dont les statistiques de Pearson et Kullback sont des cas particuliers.

5.1.1 La famille des statistiques de divergence de puissance

La famille de statistiques de divergence de puissance a été introduite puis étudiée par Read et Cressie dans [26] et [70]. Elle permet de généraliser les statistiques connues et de mieux traiter le cas des faibles effectifs n.

Les statistiques de cette famille, paramétrée par λ, sont définies par :

$$RC^\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{r=1}^{R} O_r \left[\left(\frac{O_r}{E_r} \right)^\lambda - 1 \right], \quad \lambda \in \mathbb{R},$$

où les O_r sont les effectifs observés et E_r les effectifs théoriques sous $\mathcal{H}_0$, éventuellement estimés.

Avec $E_r = np_r^{*0}$ et $O_r = N_r$, nous obtenons :

$$RC^\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{r=1}^{R} N_r \left[\left(\frac{N_r}{np_r^{*0}} \right)^\lambda - 1 \right], \quad \lambda \in \mathbb{R}.$$

Plus généralement, pour deux probabilités p et p' quelconques de $]0,1[^R$ et un échantillon de taille n :

$$RC_p^\lambda(p') = \frac{2n}{\lambda(\lambda+1)} \sum_{r=1}^{R} p'_r \left[\left(\frac{p'_r}{p_r} \right)^\lambda - 1 \right], \quad \lambda \in \mathbb{R}.$$

Nous reconnaissons en RC^λ la statistique $RC_{p^{*0}}^\lambda(p^*)$. Cette statistique permet de tester l'adéquation de l'échantillon à la loi $\mathcal{M}(n; p^0)$. Elle est ainsi souvent appelée « distance », mais ne vérifie pas la propriété de symétrie d'une distance d'un espace métrique. C'est pour cette raison qu'il est préférable de la nommer **mesure de divergence**.

Read et Cressie ont montré, sous certaines conditions de régularité notées (B) et énoncées par Birch dans [14], que les statistiques RC^λ sont toutes équivalentes asymptotiquement, et que leur loi limite est une loi du khi-deux à $R-s-1$ degrés de liberté.

Théorème 4. *Pour tout λ, sous les conditions de régularité (B) et sous $\mathcal{H}_0$, la statistique $RC_{p^{*0}}^\lambda(p^*)$ converge en loi lorsque n tend vers $+\infty$, vers une variable aléatoire suivant une loi du khi-deux à $R-s-1$ degrés de liberté. Nous notons cela :*

$$\forall \lambda \in \mathbb{R}, \quad \lim_{n \to +\infty} \mathcal{L}_{\mathcal{H}_0}(RC_{p^{*0}}^\lambda(p^*)) = \chi^2_{R-s-1}.$$

Nous introduisons à présent les statistiques de Pearson et de Kullback, et montrons qu'elles sont des cas particuliers de la famille de statistiques de divergence de puissance.

5.1.2 Statistique du khi-deux de Pearson

Historiquement, le premier test d'indépendance des caractères d'une table de contingence est le test dit « du khi-deux de Pearson », introduit en 1900 dans [64]. Pour une table à R cases, la statistique générale appelée statistique de Pearson ou du khi-deux est :

$$Q = \sum_{r=1}^{R} \frac{(O_r - E_r)^2}{E_r}.$$

Cette formule donne, avec l'estimation du maximum de vraisemblance pour E_r et les observations $O_r = N_r$:

$$Q = \sum_{r=1}^{R} \frac{(N_r - np_r^{*0})^2}{np_r^{*0}}.$$

De façon générale, nous noterons pour deux lois de probabilité p et p' quelconques de $]0,1[^R$ et un échantillon de taille n :

$$Q_p(p') = n\sum_{r=1}^{R} \frac{(p'_r - p_r)^2}{p_r} = \left(n\sum_{r=1}^{R} \frac{(p'_r)^2}{p_r} \right) - n.$$

Ainsi, la quantité Q que nous avons considérée est $Q_{p^{*0}}(p^*)$.

Théorème 5. *Sous certaines conditions de régularité et sous $\mathcal{H}_0$, la statistique $Q_{p^{*0}}(p^*)$ converge en loi lorsque n tend vers $+\infty$ vers une variable aléatoire suivant une loi du khi-deux à $R - s - 1$ degrés de liberté :*

$$\lim_{n\to+\infty} \mathcal{L}_{\mathcal{H}_0}(Q_{p^{*0}}(p^*)) = \chi^2_{R-s-1}.$$

Une preuve utilisant des conditions de régularité plus fortes que (B) est donnée dans [25]. Une autre preuve est donnée indirectement par la proposition suivante à travers la famille des statistiques de divergence de puissance et en utilisant le résultat de convergence sur cette famille. En effet :

Proposition 3. *La statistique Q correspond à la statistique RC^1.*

Démonstration.

$$\begin{aligned}
RC^1 &= \sum_{r=1}^{R} O_r \left[\left(\frac{O_r}{E_r} \right) - 1 \right] = \sum_{r=1}^{R} (O_r - E_r + E_r) \left(\frac{O_r - E_r}{E_r} \right), \\
&= \sum_{r=1}^{R} \frac{(O_r - E_r)^2}{E_r} + \sum_{r=1}^{R} (O_r - E_r), \\
&= \sum_{r=1}^{R} \frac{(O_r - E_r)^2}{E_r}, \quad \text{car} \quad \sum_{r=1}^{R} (O_r - E_r) = 0, \\
&= Q.
\end{aligned}$$

□

Une autre estimation de θ intéressante en dehors de l'estimation du maximum de vraisemblance est celle du khi-deux minimum, c'est-à-dire la valeur θ_{min} de θ qui

minimise la statistique $Q_{p^0(\theta)}(p^*)$. Avec un tel choix, nous obtenons la **statistique du khi-deux minimum de Cràmer**, définie dans [64] :

$$Q_{p^0(\theta_{min})}(p^*) = \min_{\theta\in\Theta} \sum_{r=1}^{R} \frac{(N_r - np_r^0(\theta))^2}{np_r^0(\theta)}.$$

Cette statistique a le même comportement asymptotique sous $\mathcal{H}_0$ que $Q_{p^{*0}}(p^*)$, comme prouvé dans [23], mais son étude est plus complexe parce que θ_{min} n'a pas de forme explicite. Calculer le minimum est donc une procédure numérique longue et difficile, c'est pourquoi elle est peu utilisée en pratique.

5.1.3 Statistique de Kullback

Une statistique de test usuelle dans le cadre des tests d'adéquation à une loi multinomiale est la statistique de Kullback, définie par Good dans [40], et étudiée en détails par Kullback dans [53] :

$$G = 2\sum_{r=1}^{R} O_r \ln\left(\frac{O_r}{E_r}\right) = 2\sum_{r=1}^{R} N_r \ln\left(\frac{N_r}{np_r^{*0}}\right). \tag{5.1.1}$$

Elle est également appelée statistique du minimum d'information discriminante, ou encore **MDIS** pour Minimum Discrimination Information Statistic.

De façon générale, nous notons pour deux lois de probabilité p et p' quelconques de $]0,1[^R$ et un échantillon de taille n :

$$G_p(p') = 2n\sum_{r=1}^{R} p'_r \ln\frac{p'_r}{p_r}, \tag{5.1.2}$$

ainsi la quantité G que nous avons considérée est $G_{p^{*0}}(p^*)$ qui dépend de n.

Notons que les équations du maximum de vraisemblance pour l'estimation de θ sont également celles qui résolvent le problème de la minimisation de la statistique de Kullback en θ, comme expliqué dans [53].

Cette statistique tire ses origines de la théorie de l'information développée par Kullback. Notons que les statistiques de Pearson et Kullback sont respectivement proportionnelles aux distances de Pearson et Kullback-Leibler. L'une des caractéristiques intéressantes de la statistique G démontrée par Kullback dans [53], est qu'elle est fortement liée à l'hypothèse nulle d'indépendance posée pour le test : elle peut se décomposer en différentes sommes partielles nommées composantes additives, possédant chacune un degré de liberté spécifique et pouvant être testées individuellement. Celles-ci sont ensuite additionnées selon que nous nous intéressons à une indépendance mutuelle, partielle, conditionnelle, ou à des interactions entre caractères, comme exposé dans [53, 71]. Cette démarche est proche de celle induite par les modèles log-linéaires.

Proposition 4. *La statistique G correspond à la limite de RC^λ lorsque λ tend vers 0, que nous notons RC^0 par continuité.*

Démonstration.
Développons l'écriture du terme général de RC^λ :

$$\begin{aligned}\frac{2}{\lambda(\lambda+1)}\sum_{r=1}^{R} O_r\left[\left(\frac{O_r}{E_r}\right)^{\lambda}-1\right] &= \frac{2}{\lambda(\lambda+1)}\sum_{r=1}^{R} O_r\left[\exp\left(\lambda\ln\left(\frac{O_r}{E_r}\right)\right)-1\right],\\ &= \frac{2}{\lambda(\lambda+1)}\sum_{r=1}^{R} O_r\left(\lambda\ln\left(\frac{O_r}{E_r}\right)+o(\lambda)\right).\end{aligned}$$

Lorsque nous faisons tendre λ vers 0 dans l'expression précédente, nous obtenons :

$$RC^0 = 2\sum_{r=1}^{R} O_r\ln\left(\frac{O_r}{E_r}\right) = G.$$

□

La convergence en loi sous $\mathcal{H}_0$ de la statistique de Kullback vers une loi du khi-deux est une conséquence de son appartenance à la famille des divergences de puissance. De plus, Agresti montre directement dans [2] que G est asymptotiquement équivalente à Q sous l'hypothèse nulle, en utilisant un développement limité et la convergence des estimateurs du maximum de vraisemblance.

5.1.4 Validité des tests et tables creuses

Les statistiques de la famille de divergence de puissance ainsi que la statistique $Q_{p^0(\theta_{min})}(p^*)$ ont toutes le même comportement asymptotique. Leur convergence en loi vers une loi du khi-deux lorsque $n \to +\infty$ est une conséquence du théorème de la limite centrale. Ce dernier impose des conditions d'utilisation pratique, parce que l'approximation n'est valable que pour une taille d'échantillon et des effectifs théoriques E_r « assez grands », ou « pas trop petits ». Lorsque ces conditions ne sont pas respectées, il est conseillé de regrouper des catégories voisines et cohérentes. Cela n'est cependant pas toujours possible, comme dans l'exemple des données génomiques où de tels regroupements n'ont aucun sens parce que les catégories ne peuvent être ordonnées.

Un certain nombre de résultats empiriques existent quant aux conditions de validité du test. Pour cette raison, nous supposons en général que les conditions suivantes sont respectées avant d'effectuer un test de Pearson ou de Kullback :

$$n \geqslant 30 \text{ et } E_r = np_r^{*0} \geqslant 5,\ \forall r \in \{1,\ldots,R\}.$$

Conover donne quant à lui dans [24] les consignes empiriques suivantes moins strictes, mais arbitraires, issues de [23]. Il autorise le test lorsque :

$$E_r \geqslant 0.5 \quad , \forall r \in \{1, \ldots, R\},$$

ou lorsque au moins la moitié des E_r vérifie $E_r \geqslant 1$, et le déconseille sinon. Dans [2], Agresti propose de faire confiance à l'approximation asymptotique du test lorsque $n/R \geqslant 5$, ou de façon moins restrictive (mais assez floue) quand $n/R > 1$ avec I, J ou K grand et si la loi théorique ne comprend pas à la fois des effectifs E_r très grands et très petits. D'autres études empiriques aboutissant à de telles conditions figurent par exemple dans [51, 88].

Sous ces conditions, il est alors légitime de se demander laquelle des statistiques proposées est la plus appropriée. En effet, même si $Q_{p^{*0}}(p^*)$ et $G_{p^{*0}}(p^*)$ ont la même distribution asymptotique et des valeurs similaires lorsque les conditions de convergence sont respectées, il n'en est pas de même dans le cas contraire et leurs valeurs pour des tables de contingence creuses peuvent différer de beaucoup. Des simulations montrent par exemple dans [3] que la convergence est plus rapide pour $Q_{p^{*0}}(p^*)$ que pour $G_{p^{*0}}(p^*)$. De par son origine dans la théorie de l'information, il est usuel de préférer $G_{p^{*0}}(p^*)$.

Plus généralement, le choix entre les statistiques de la famille peut s'effectuer sur la base de critères variés. Par un calcul de moments, Read et Cressie recommandent dans [26] de choisir $\lambda = 2/3$ pour obtenir une statistique qui serait un compromis intéressant entre $Q = Q_{p^{*0}}(p^*)$ et $G = G_{p^{*0}}(p^*)$ lorsque $\min_r E_r \geqslant 1$ et $n \geqslant 10$. Ils comparent dans [70] la statistique $RC^{2/3} = RC^{2/3}_{p^{*0}}(p^*)$ à Q et G à l'aide de risques empiriques de première espèce et de puissances, calculés pour des tables de contingence de dimension deux, en faisant varier le nombre de catégories, la taille n de l'échantillon et la forme de l'hypothèse nulle.

Cependant, ces conditions, même peu restrictives, ne sont pas vérifiées pour des tables creuses, d'où la nécessité de trouver une alternative à cette famille.

5.1.5 Test exact de Fisher

Il existe un test exact donné par Fisher dans [33] qui est valable sans restrictions. Une description détaillée de ce test figure dans [24]. Il est fondé sur le calcul exact des probabilités en fixant les marges, et en utilisant la loi hypergéométrique. Le degré de significativité de la table étudiée est égal à la somme des probabilités de toutes les tables qui sont moins probables qu'elle. Il faut donc énumérer un grand nombre de tables et ceci devient numériquement très coûteux lorsque le nombre de catégories augmente.

Lorsque le test exact n'est pas réalisable, il est possible d'employer des méthodes de Monte Carlo, comme suggéré au chapitre 7 de [2]. Nous allons nous concentrer

sur un autre moyen de traiter le cas des tables creuses en corrigeant les statistiques de test en fonction du nombre de zéros.

5.2 Correction de Ku pour la statistique de Kullback

Pensant qu'il fallait diminuer la statistique de Kullback dans le cas de tables avec des zéros, Ku introduit dans [52] une correction soustractive de $G_{p^{*0}}(p^*)$ d'une valeur de 1 pour un seul zéro dans la table. Il en suggère une extension, sans la justifier, au cas d'un nombre quelconque de zéros. Ku montre l'intérêt de sa correction sur l'échantillon de taille 10 donné par Pearson dans [61] dans le cas d'un seul zéro. Cette démarche ne nous paraît pas fondée. D'une part cet échantillon est trop petit pour tester un tel résultat asymptotique, d'autre part la correction de Ku ne peut s'étendre, comme il l'affirme, au cas de plusieurs zéros sans modifier la démonstration en conséquence.

Supposons qu'après renumérotation, nous ayons $n_1 = n_2 = \cdots = n_c = 0$ et que pour tout j de $\{c+1, \ldots, R\}, n_j \geqslant 1$. Supposons également qu'il y ait au moins une case non nulle, donc que $R - c \geqslant 1$.

5.2.1 Statistique de Kullback et tables creuses

L'argument de Ku pour sa correction de la statistique de Kullback, et pour la rapprocher de celle de Pearson en présence d'au moins un zéro est le suivant. Lorsque la table est creuse, les tests d'hypothèses utilisant $G_{p^{*0}}(p^*)$ ne sont plus fiables parce que cette statistique prend des valeurs trop élevées par rapport à $Q_{p^{*0}}(p^*)$. L'estimateur du maximum de vraisemblance pour p vaut :

$$p^* = (p_1^*, \ldots, p_R^*) = \left(\frac{n_1}{n}, \ldots, \frac{n_R}{n}\right) = \left(0, \ldots, 0, \frac{n_{c+1}}{n}, \ldots, \frac{n_R}{n}\right),$$

sous-estimant les p_i pour $i \in \{1, \ldots, c\}$ et sur-estimant les p_j pour $j \in \{c+1, \ldots, R\}$. Selon Ku, la présence de zéros dans la table tend à augmenter artificiellement certains termes de la statistique $G_{p^0}(p^*)$ à cause la proposition suivante :

Proposition 5. *Pour tous $a, b > 0$ tels que $a < 2b$, $b < 2a$ et tels que les quantités $1/a$ et $1/b$ soient majorées, nous avons :*

$$\ln\left(\frac{a}{b}\right) = \frac{a^2 - b^2}{2ab} + o(a - b).$$

De plus, si $a = b$ alors l'égalité est réalisée.

Démonstration.
Pour a et b vérifiant les hypothèses de l'énoncé, nous pouvons écrire les développements limités suivants à l'ordre 1 :

$$\ln\left(\frac{a}{b}\right) = \ln\left(1+\left(\frac{a-b}{b}\right)\right) = \frac{a-b}{b} + o\left(\frac{a-b}{b}\right),$$
$$\ln\left(\frac{a}{b}\right) = -\ln\left(\frac{b}{a}\right) = -\ln\left(1+\left(\frac{b-a}{a}\right)\right) = \frac{a-b}{a} + o\left(\frac{a-b}{a}\right).$$

Nous en déduisons par sommation que :

$$\begin{aligned}\ln\left(\frac{a}{b}\right) &= \frac{1}{2}\left[\frac{a-b}{a}+\frac{a-b}{b}\right] + o\left(\frac{a-b}{b}\right) + o\left(\frac{a-b}{a}\right),\\ &= \frac{a^2-b^2}{2ab} + o(a-b).\end{aligned}$$

En effet, comme $1/a$ et $1/b$ sont majorés, une quantité $o((a-b)/b)$ ou $o((a-b)/a)$ est également $o(a-b)$. □

Pour r dans $\{1,\ldots,R\}$, posons $a = n_r/n$ et $b = p_r^{*0}$. Lorsque n/n_r et $1/p_r^{*0}$ sont majorés, nous déduisons de cette proposition que :

$$\ln\left(\frac{n_r}{np_r^{*0}}\right) = \frac{n_r^2-(np_r^{*0})^2}{2n_r np_r^{*0}} + o\left(\frac{n_r}{n}-p_r^{*0}\right).$$

Ku applique l'approximation du premier ordre :

$$2n_r \ln\left(\frac{n_r}{np_r^{*0}}\right) \simeq \frac{n_r^2-(np_r^{*0})^2}{np_r^{*0}}$$

aux valeurs de r pour lesquelles $n_r = 0$. Le membre de gauche est alors nul, tandis que le membre de droite est négatif. Les sommes sur r de chaque membre correspondent respectivement à la statistique $G_{p^{*0}}(p^*)$ pour le membre de gauche, et à $Q_{p^{*0}}(p^*)$ pour le membre de droite. Selon lui, cela implique que dans le cas de zéros, G est toujours plus grande que Q, et qu'il faut donc corriger la statistique de Kullback pour la ramener au niveau de la statistique de Pearson.

Cependant, si les termes de la statistique de Kullback correspondant à des zéros dans la table sont supérieurs à ceux de la statistique de Pearson, ce phénomène ne vaut pas pour les autres termes, donc pas pour la statistique générale. Un phénomène de report d'effectifs implique que des cases soient plus chargées pour compenser les zéros. Il n'est donc pas possible d'ordonner les statistiques comme le pense Ku.

De plus, Ku applique la proposition dans un cas où l'approximation n'est pas valable. En effet, lorsque $n_r = 0$, le quotient n/n_r n'est pas majoré, et le rapport

n_r/np_r^{*0} est nul. Le raisonnement de Ku ne s'applique donc pas dans ce cas, même si nous remarquons bien un écart dans les valeurs des deux statistiques lorsque le nombre de zéros augmente, ainsi qu'une déviation par rapport à la valeur critique de la loi limite, malgré des valeurs élevées de n. Pour le voir, nous procédons à des simulations selon différentes lois multinomiales.

5.2.2 Simulations

Nous allons voir qu'il est en réalité nécessaire de corriger à la fois Q et G puisque leurs quantiles s'éloignent tous les deux de la valeur asymptotique théorique représentée par le quantile d'ordre $1-\alpha$ de la loi du khi-deux au nombre de degrés de liberté correspondant à $\mathcal{H}_0$.

Nous simulons 1 000 vecteurs d'effectif total $n = 400$ selon chacune des 6 lois multinomiales à $R = 100$ catégories, de probabilités f_1 à f_6 données dans la table 5.1. Nous y indiquons également le nombre $|\{r : E_r < 0.5\}|$ de catégories pour lesquelles les effectifs théoriques E_r sont strictement inférieurs à 0.5. Cette table présente un nombre croissant de petits effectifs théoriques car les statistiques sont construites avec un nombre croissant de catégories de très petite probabilité.

$\mathcal{H}_0$	$\lvert\{r :\ E_r < 0.5\}\rvert$	Vecteur de probabilité
f_1	0	$(0.01, \ldots, 0.01)$
f_2	20	$(\underbrace{0.001, \ldots, 0.001}_{20\text{ fois}}, \underbrace{0.01, \ldots, 0.01}_{60\text{ fois}}, \underbrace{0.019, \ldots, 0.019}_{20\text{ fois}})$
f_3	30	$(\underbrace{0.003, \ldots, 0.003}_{30\text{ fois}}, \underbrace{0.99/70, \ldots, 0.99/70}_{70\text{ fois}})$
f_4	40	$(\underbrace{0.004, \ldots, 0.004}_{40\text{ fois}}, \underbrace{0.99/60, \ldots, 0.99/60}_{60\text{ fois}})$
f_5	91	$(\underbrace{0.1, \ldots, 0.1}_{9\text{ fois}}, \underbrace{0.1/91, \ldots, 0.1/91}_{91\text{ fois}})$
f_6	88	$(0.5, 1/3, \underbrace{1/70, \ldots, 1/70}_{10\text{ fois}}, \underbrace{1/3696, \ldots, 1/3696}_{88\text{ fois}})$

Table 5.1 – Probabilités multinomiales f_1 à f_6.

c	0	1	2	3	4	5	6	7
$\lvert c\rvert$	156	299	297	157	68	15	7	1

TABLE 5.2 – Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_1, selon c.

c	6	7	8	9	10	11	12	13	14
$\lvert c\rvert$	1	2	6	6	27	69	80	121	170
c	15	16	17	18	19	20	21	22	23
$\lvert c\rvert$	181	157	91	55	22	8	2	1	1

TABLE 5.3 – Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_2, selon c.

c	20	21	22	23	24	25	26	27	28	29	30	31	32
$\lvert c\rvert$	2	3	12	45	57	161	210	190	188	94	33	4	1

TABLE 5.4 – Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_3, selon c.

c	31	32	33	34	35	36	37	38	39	40	41
$\lvert c\rvert$	5	30	42	84	168	211	212	153	70	19	6

TABLE 5.5 – Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_4, selon c.

c	45	47	48	49	50	51	52	53	54	55	56	57	58	59
$\lvert c\rvert$	1	5	7	2	14	18	28	47	57	56	75	95	89	84
c	60	61	62	63	64	65	66	67	68	69	70	71	73	75
$\lvert c\rvert$	86	69	62	53	48	32	26	15	16	7	3	3	1	1

TABLE 5.6 – Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_5, selon c.

c	70	71	72	73	74	75	76	77	78
$\lvert c\rvert$	3	4	9	23	33	31	71	91	137
c	79	80	81	82	83	84	85	86	
$\lvert c\rvert$	131	157	105	99	68	24	11	3	

TABLE 5.7 – Effectifs $|c|$ pour le calcul des quantiles d'ordre 0.95 pour f_6, selon c.

Les graphes de la figure 5.1 représentent l'évolution des quantiles des statistiques Q et G en fonction du nombre de zéros, déterminé par la probabilité de la loi multinomiale utilisée pour générer les observations. Le seuil choisi est $\alpha = 0.05$, et le quantile de la loi du khi-deux à $R - 1 = 99$ degrés de liberté correspondant vaut $\chi^2_{0.95,99} = 123.22$. Pour chaque hypothèse et chaque valeur de c, le nombre de valeurs $|c|$ utilisées parmi les 1 000 valeurs simulées pour calculer le quantile est donné dans les tables 5.2 à 5.7.

Ils montrent que Q a tendance à exploser lorsque c augmente. En effet, les valeurs du quantile de la statistique Q augmentent avec le nombre de zéros, alors que les quantiles de la statistique G restent stables. Ceci peut s'expliquer par le fait que Q converge plus vite que G vers la loi asymptotique du khi-deux. L'espérance d'une telle loi étant égale à son nombre de degrés de liberté, si celui-ci est grand, Q se rapproche rapidement d'une valeur élevée. Nous n'observons rien de tel pour la statistique G. Alors que Q est majoritairement au-dessus du seuil critique $\chi^2_{0.95,99}$, la statistique G passe en-dessous du seuil à partir de f_3. Ces résultats viennent contredire Ku, et suggèrent de corriger surtout la statistique Q.

Nous remarquons en comparant les deux premiers graphes que pour des vecteurs comportant un nombre identique de zéros, mais générés suivant deux lois multinomiales différentes, le comportement des statistiques diffère. Ce phénomène provient certainement des petits échantillons sur lesquels ont été calculés les quantiles. En effet, nous pouvons voir dans les tables 5.2 à 5.7 que certains quantiles n'ont été calculés que sur une observation. Nous ne devons donc nous intéresser qu'à la partie centrale de chaque graphe, pour laquelle les effectifs sont assez importants pour que le calcul du quantile ait un sens.

Nous calculons pour Q et G les risques empiriques de première espèce pour les hypothèses de probabilités f_1 à f_6, pour plusieurs valeurs du seuil. Nous donnons les résultats dans le tableau suivant, en fonction du mode des zéros $mode(c)$, c'est-à-dire la valeur du nombre de zéros réalisée par un maximum de tables.

Nous observons que les risques empiriques pour Q s'éloignent de α par valeurs supérieures lorsque c augmente. Ceci contribue à mettre en doute les conclusions de Ku données au paragraphe 5.2.1 selon lesquelles Q est la référence pour l'ajustement de G, et nous conforte dans notre volonté de corriger à la fois Q et G.

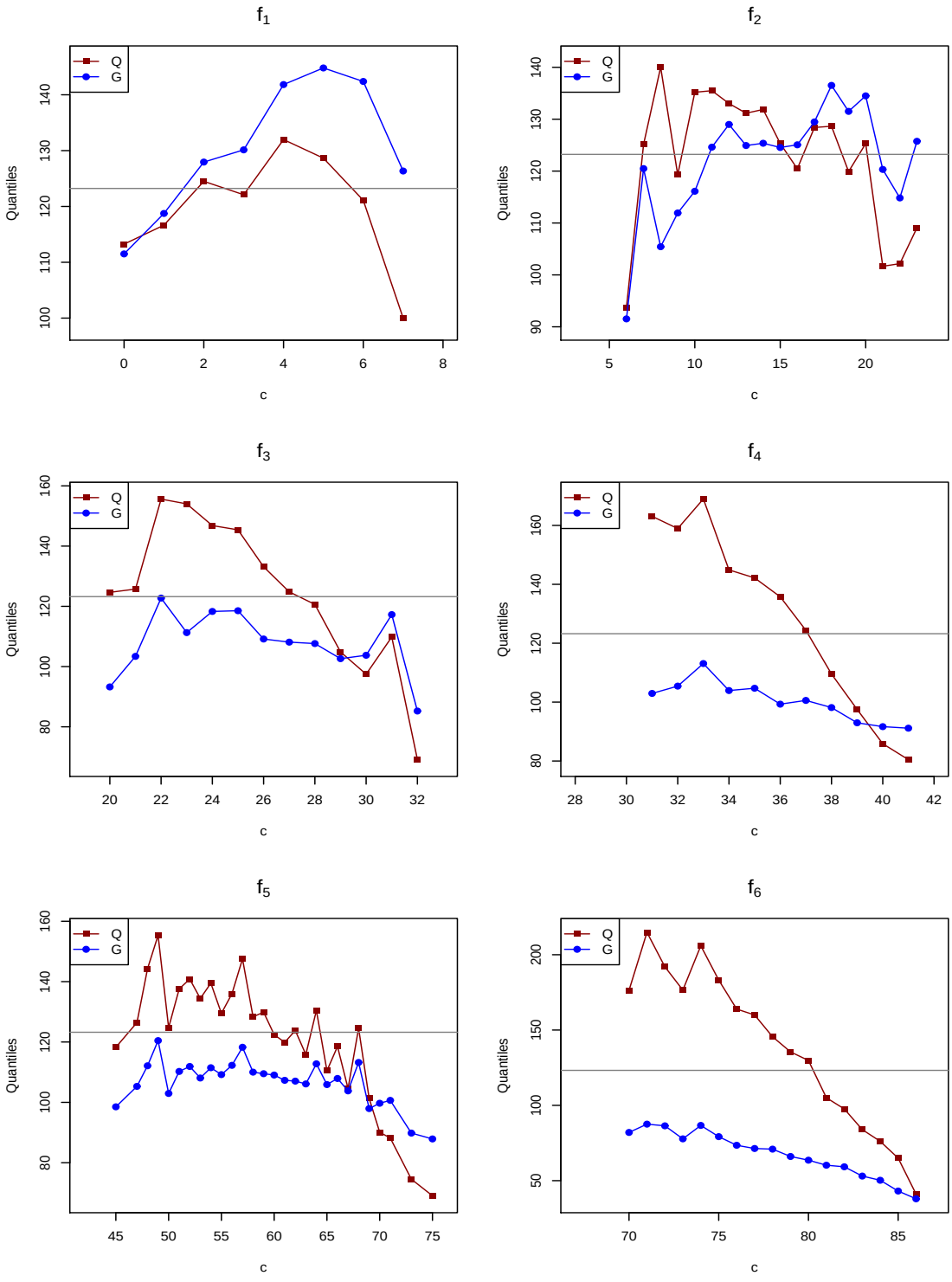


FIGURE 5.1 – Quantiles d'ordre 0.95 de Q et G en fonction de c, sous les hypothèses nulles f_1 à f_6, pour 1 000 échantillons de $n = 400$ individus répartis dans $R = 100$ catégories. La ligne horizontale représente le seuil critique $\chi^2_{0.95,99}$.

	$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.1$		
$\mathcal{H}_0$	$mode(c)$	Q	G	$mode(c)$	Q	G	$mode(c)$	Q	G
f_1	1	0	0	1	0.027	0.027	1	0.064	0.070
f_2	14	0	0.005	15	0.061	0.061	15	0.082	0.123
f_3	27	0.029	0	26	0.086	0.009	26	0.141	0.009
f_4	36	0.051	0	37	0.061	0	37	0.059	0
f_5	59	0.067	0	57	0.168	0.011	58	0.151	0
f_6	79	0.034	0	80	0.089	0	79	0.128	0

Table 5.8 – Comparaison des risques empiriques de première espèce pour les statistiques Q et G, pour 1 000 échantillons de $n = 400$ individus répartis dans $R = 100$ catégories, aux seuils 0.01, 0.05, 0.1, pour les lois multinomiales de probabilités f_1 à f_6 et le mode du nombre de zéros correspondant.

5.2.3 Cas d'un zéro unique

Plaçons-nous dans le cas d'un vecteur comportant seulement un zéro, placé en première position. Nous donnons ici la justification de Ku dans [52] pour diminuer la statistique G de la valeur 1 lorsqu'une case est nulle. Pour le zéro, nous remplaçons la valeur nulle de l'estimateur du maximum de vraisemblance par une valeur non nulle. Soit p le véritable paramètre de la loi multinomiale.

Nous supposons que $0 < p_1 < 1$, parce que nous considérons un zéro aléatoire et non structural, et que la table comporte au moins deux catégories. Comme $x = (n_1, \ldots, n_R)$ est une réalisation de $X \sim \mathcal{M}(n; p)$, nous avons par la propriété 2 du chapitre 2 $N_1 \sim \mathcal{B}(n; p_1)$, et donc le principe selon lequel nous avons observé la valeur la plus probable implique :

$$\begin{aligned} \mathbb{P}(N_1 = 0) > \mathbb{P}(N_1 = 1) &\iff (1 - p_1)^n > np_1(1 - p_1)^{n-1}, \\ &\iff 1 - p_1 > np_1, \\ &\iff p_1 < \frac{1}{n+1}. \end{aligned}$$

Nous nommons ce type de raisonnement **principe de Fisher**. Pour la première case, nous choisissons donc comme estimateur de p_1 la moyenne de la loi uniforme sur le segment $]0, \frac{1}{n+1}[$, soit

$$\hat{p}_1 = \frac{1}{2(n+1)}.$$

Les estimateurs des cases non nulles sont les estimateurs du maximum de vraisemblance correspondants, diminués d'une constante telle que les $\hat{p}_r$ somment à 1. Nous obtenons alors l'estimateur suivant :

$$\begin{cases} \hat{p}_1^{\mathrm{Ku}} & = \dfrac{1}{2(n+1)}, \\ \hat{p}_j^{\mathrm{Ku}} & = \dfrac{n_j}{n} - \dfrac{1}{2(R-1)(n+1)}, \qquad \forall j \in \{2, \ldots, R\}. \end{cases} \tag{5.2.1}$$

Nous obtenons la statistique de Kullback corrigée :

$$\begin{aligned} G^{\mathrm{Ku}} &= 2\sum_{j=1}^{R} n_j \ln\left(\frac{n\hat{p}_j^{\mathrm{Ku}}}{np_j^{*0}}\right), \\ &= 2\sum_{j=2}^{R} n_j \ln\left(\frac{1}{np_j^{*0}}\left(n_j - \frac{n}{2(R-1)(n+1)}\right)\right), \\ &= 2\sum_{j=2}^{R} n_j \ln\frac{n_j}{np_j^0} + 2\sum_{j=2}^{R} n_j \ln\left(1 - \frac{n}{2(R-1)(n+1)n_j}\right). \end{aligned}$$

En reconnaissant le premier terme comme la statistique de Kullback et à l'aide d'un développement limité, Ku obtient :

$$G^{\mathrm{Ku}} = G_{p^{*0}}(p^*) - \frac{n}{n+1} + o(1) = G_{p^{*0}}(p^*) - 1 + o(1).$$

Ainsi, la correction proposée par Ku consiste à soustraire la valeur 1 à la statistique de Kullback. Nous remarquons que Ku n'a pas exactement remplacé p^* par $\hat{p}$ dans $G_{p^0}(p^*)$, mais qu'il l'a fait uniquement dans le logarithme de l'expression. Ce choix est peut-être dû à une volonté de simplicité, ou à une confusion liée à la forme de la statistique qui s'écrit selon (5.1.1) dans le cas particulier du modèle multinomial, et selon (5.1.2) dans le cas général qu'il est préférable de considérer. Nous verrons plus tard comment être plus cohérents que Ku, tout en proposant une solution facilement implémentable.

L'exemple que Ku utilise pour illustrer sa correction s'inspire de celui de Neyman et Pearson dans [61]. À la fin de leur article dans lequel ils étudient la pertinence de l'approximation de $Q_{p^0}(p^*)$ par une loi du khi-deux sur de petits échantillons à l'aide de degrés de significativité, ces derniers avouent ne pas connaître l'impact de fréquences théoriques très petites sur cette même loi. Dans leur exemple de loi multinomiale $\mathcal{M}(10; (0.3,\ 0.5,\ 0.2))$, ils sont agréablement surpris par le peu de déviation que cela engendre. En revanche, il est légitime de se demander si un échantillon de seulement 10 individus peut être considéré comme assez grand pour comparer les différents tests asymptotiques, comme le fait Ku. Ce dernier propose ensuite une version de cette correction généralisée à un nombre quelconque de zéros.

5.2.4 Cas de zéros multiples

Nous allons voir en quoi la correction proposée par Ku dans le cas de plusieurs zéros n'est pas adaptée. Nous explicitons le résultat suggéré par Ku dans le cas de plusieurs zéros, puis nous étudions sa validité sur des simulations. Nous raisonnons de la même manière que dans le cas d'un zéro unique, et supposons cette fois :

$$n \geqslant 2,\ c \geqslant 2,\ R \geqslant 3,\ R - c \geqslant 1.$$

Soit p le paramètre de la loi multinomiale. Comme $x = (n_1, \ldots, n_R)$ est une réalisation de $X \sim \mathcal{M}(n; p)$, la variable N_r suit une loi $\mathcal{B}(n; p_r)$ par la propriété 2 du chapitre 2, et donc nous avons pour tout i dans $\{1, \ldots, c\}$ selon le principe de Fisher :

$$\mathbb{P}(N_i = 0) > \mathbb{P}(N_i = 1) \iff p_i < \frac{1}{n+1}.$$

Comme dans le cas $c = 1$, nous choisissons alors, pour $c \in \{2, \ldots, R-1\}$:

$$\begin{cases} \hat{p}_i^{\mathrm{Ku}} &= \dfrac{1}{2(n+1)}, & \forall i \in \{1, \ldots, c\}, \\ \hat{p}_j^{\mathrm{Ku}} &= \dfrac{n_j}{n} - \dfrac{c}{2(R-c)(n+1)}, & \forall j \in \{c+1, \ldots, R\}. \end{cases} \tag{5.2.2}$$

Notons que pour j dans $\{c+1, \ldots, R\}$,

$$\hat{p}_j^{\mathrm{Ku}} = \frac{2n_j(R-c)(n+1) - cn}{2n(R-c)(n+1)}.$$

Prenons par exemple $R = 7$, $n = 10$ et $x = (0, 0, 0, 0, 0, 1, 9)$. Alors le numérateur de $\hat{p}_6^{\mathrm{Ku}}$ est négatif, donc $\hat{p}_6^{\mathrm{Ku}}$ l'est aussi. Le problème vient du fait que Ku ne considère pas la bonne expression de la statistique, et ne manipule donc plus des lois de probabilité. Il opère sa correction sur la version de G spécifique au maximum de vraisemblance, alors qu'il devrait revenir à son écriture sous la forme d'une distance, qui est valable pour tous les choix de p et p'.

Nous avons ainsi montré que le choix d'estimateurs de Ku consistant à généraliser l'estimateur défini pour un seul zéro n'est pas valable dans le cas de plusieurs zéros. En revanche, nous allons voir qu'il est possible d'adapter sa démarche pour obtenir la correction qui consiste à diminuer la statistique de la valeur c correspondant au nombre de zéros. Nous proposons donc d'introduire un paramètre $h \in \mathbb{R}$ qu'il nous faut choisir, et de définir :

Définition 6.

$$\begin{cases} \hat{p}_i^{h} &= \dfrac{1}{2(n+h)}, & \forall i \in \{1, \ldots, c\}, \\ \hat{p}_j^{h} &= \dfrac{n_j}{n} - \dfrac{c}{2(R-c)(n+h)}, & \forall j \in \{c+1, \ldots, R\}. \end{cases} \tag{5.2.3}$$

Notons que le choix $h = 1$ ramène aux calculs de Ku avec un seul zéro.

Proposition 6. *Le vecteur $\hat{p}^h$ donné par* (5.2.3) *définit une loi de probabilité si h est choisi tel que :*

$$h > \max\left\{\frac{1}{2} - n, n\left(\frac{c}{2\underline{n}(R-c)} - 1\right)\right\}, \text{ où } \quad \underline{n} = \min_{j\in\{c+1,\dots,R\}}\{n_j\}. \tag{5.2.4}$$

Démonstration.
Calculons tout d'abord :

$$\sum_{k=1}^{R}\hat{p}_k^h = \sum_{i=1}^{c}\hat{p}_i^h + \sum_{j=c+1}^{R}\hat{p}_j^h = \sum_{i=1}^{c}\frac{1}{2(n+h)} + \sum_{j=c+1}^{R}\left(\frac{n_j}{n} - \frac{c}{2(R-c)(n+h)}\right) = 1.$$

Par la condition (5.2.4), $h > 1/2 - n$. Nous avons alors pour tout $i \in \{1,\dots,c\}$ l'inégalité $0 < \hat{p}_i^h < 1$. En effet, $h > -n$ donne $\hat{p}_i^h > 0$, et $h > 1/2 - n$ implique directement $\hat{p}_i^h < 1$.
Soit $j \in \{c+1,\dots,R\}$. Nous avons :

$$\frac{c}{2(R-c)(n+h)} > 0,$$

puis

$$\hat{p}_j^h = \frac{n_j}{n} - \frac{c}{2(R-c)(n+h)} < \frac{n_j}{n} \leqslant 1.$$

Il reste à voir que $\hat{p}_j^h > 0$. Comme :

$$\hat{p}_j^h = \frac{n_j}{n} - \frac{c}{2(R-c)(n+h)} = \frac{2n_j(R-c)(n+h) - cn}{2n(R-c)(n+h)},$$

il suffit de montrer que $2n_j(R-c)(n+h) - cn > 0$. La condition (5.2.4) donne en particulier $h > \frac{cn}{2n_j(R-c)} - n$. Cette inégalité implique $2n_j(R-c)(n+h) > cn$, ce que nous voulions démontrer.

□

Écrivons à présent la statistique de Kullback corrigée G^h adaptée au cas de plusieurs zéros :

$$\begin{aligned}
G^h &= 2\sum_{i=1}^{c} n_i \ln\left(\frac{\hat{p}_i^h}{p_i^{*0}}\right) + 2\sum_{j=c+1}^{R} n_j \ln\left(\frac{\hat{p}_j^h}{p_j^{*0}}\right), \\
&= 2\sum_{j=c+1}^{R} n_j \ln\left(\frac{1}{p_j^{*0}}\left(\frac{n_j}{n} - \frac{c}{2(R-c)(n+h)}\right)\right), \\
&= 2\sum_{j=c+1}^{R} n_j \ln\left(\frac{n_j}{np_j^{*0}}\right) + 2\sum_{j=c+1}^{R} n_j \ln\left(1 - \frac{cn}{2n_j(R-c)(n+h)}\right).
\end{aligned}$$

Si nous choisissons h qui vérifie la condition (5.2.4), nous avons pour tout j dans $\{c+1,\ldots,R\}$ l'inégalité :

$$0 < \frac{cn}{2n_j(R-c)(n+h)} < 1,$$

et le développement du logarithme au voisinage de 1 à l'ordre 1 implique :

$$\begin{aligned} G^h &= 2\sum_{j=c+1}^{R} n_j \ln\left(\frac{n_j}{np_j^{*0}}\right) + 2\sum_{j=c+1}^{R} n_j\left(-\frac{cn}{2n_j(R-c)(n+h)}\right) + o(1), \\ &= G_{p^{*0}}(p^*) - \frac{cn}{n+h} + o(1). \end{aligned}$$

Notre correction s'applique pour $c = 1$, et le choix $h = 1$ ramène à la correction de Ku pour un seul zéro. En pratique, nous choisissons h le plus petit possible, ce qui revient à choisir :

$$h = \max\left\{\frac{1}{2} - n, n\left(\frac{c}{2\underline{n}(R-c)} - 1\right)\right\} + \epsilon,$$

où ϵ est une petite constante destinée à éviter les effets de bord.

5.2.5 Simulations

Les simulations réalisées pour les différentes fréquences multinomiales présentées au paragraphe 5.2.2 montrent que la correction réalisée par la statistique G^h est négligeable par rapport à G. En effet, nous avons observé des différences relatives entre G et G^h, allant de 10^{-6} à 10^{-2}. Nous remarquons en réalité que la valeur minimale de h est très grande quel que soit n, et que la correction en $cn/(n+h)$ est négligeable même lorsque n est petit.

Pour ces raisons, nous proposons dans la suite une nouvelle correction pour $G_{p^{*0}}(p^*)$, mais également pour $Q_{p^{*0}}(p^*)$, que Ku avait lui choisi de ne pas corriger. Il considère en effet cette statistique de Pearson comme une référence « robuste » aux zéros. Les simulations du paragraphe 5.2.2 nous montrent le contraire, c'est pourquoi nous corrigeons à la fois G et Q. Ces corrections tiennent bien sûr compte comme pour G^h du nombre de zéros c.

5.3 Corrections des statistiques de Pearson et Kullback

La méthode que nous proposons consiste à modifier les statistiques de Pearson et de Kullback en améliorant les corrections proposées par Ku.

5.3.1 Contexte et nécessité de nouvelles corrections

Rappelons que nous cherchons à tester dans le cadre d'échantillons multinomiaux $\mathcal{H}_0 : p = p^0$ contre $\mathcal{H}_1 : p \neq p^0$, où p^0 dépend éventuellement d'un paramètre θ de dimension s à estimer. Nous supposons que tous les zéros sont situés aux premières positions de x :

$$n_1 = \ldots = n_c = 0 < 1 \leqslant n_{c+1}, \ldots, n_R.$$

L'estimateur du maximum de vraisemblance, bien qu'imparfait comme noté au paragraphe 5.2.1, est utilisé dans les statistiques de test $Q_{p^{*0}}(p^*)$ et $G_{p^{*0}}(p^*)$. Nous proposons dans cette partie un autre estimateur de p, noté $\hat{p}$, sous la forme :

$$\begin{cases} \hat{p}_i & = a, & \forall i \in \{1, \ldots, c\}, \\ \hat{p}_j & = \dfrac{n_j}{n^b} - d, & \forall j \in \{c+1, \ldots, R\}, \end{cases}$$

où $a = a_n,\ b = b_n$ et $d = d_n$ sont des variables aléatoires dépendant de n à définir telles qu'elles compensent les sur- et sous-estimations lors de l'utilisation de p^*. Ainsi, nous les prenons strictement positives, avec $0 < b < 1$, pour que b « augmente » l'estimateur du maximum de vraisemblance n_j/n^b, que nous diminuons ensuite de la valeur d. Nous en déduisons des statistiques de test corrigées $Q_{p^{*0}}(\hat{p})$ et $G_{p^{*0}}(\hat{p})$ ayant un meilleur comportement que les statistiques non corrigées

5.3.2 Choix des corrections

Les paramètres $a,\ b$ et d doivent vérifier certaines conditions pour nous assurer que $\hat{p}$ définit bien une loi de probabilité. La sommation des probabilités $\hat{p}_r$ à 1 donne tout d'abord la relation :

$$(R-c)d = ac + n^{1-b} - 1.$$

Ainsi, d peut s'écrire simplement en fonction de a et b, ce qui nous donne l'expression suivante pour les probabilités :

Définition 7.

$$\begin{cases} \hat{p}_i^{ab} & = a, & \forall i \in \{1, \ldots, c\}, \\ \hat{p}_j^{ab} & = \dfrac{n_j}{n^b} - \dfrac{ac + n^{1-b} - 1}{R-c}, & \forall j \in \{c+1, \ldots, R\}, \end{cases} \tag{5.3.1}$$

et nous prenons $a = 0$ et $b = 1$ pour $c = 0$. Posons comme au paragraphe précédent $\underline{n} = \min_{j \in \{c+1,\ldots,R\}}\{n_j\}$, et $\overline{n} = \max_{j \in \{c+1,\ldots,R\}}\{n_j\}$. Les inégalités $0 < \hat{p}_i^{ab} < 1$, pour tout i dans $\{1, \ldots, c\}$ et $0 < \hat{p}_j^{ab} < 1$, pour tout j dans $\{c+1, \ldots, R\}$ sont respectivement équivalentes à :

$$0 < a < 1 \quad \text{et} \quad K_1(b) < a < K_2(b),$$

avec

$$K_1(b) = \frac{(\overline{n} - n^b)(R-c) + n^b - n}{cn^b}$$

et

$$K_2(b) = \frac{\underline{n}(R-c) + n^b - n}{cn^b}.$$

Ainsi, b doit être choisi tel que $K_1(b) < 1$, $K_2(b) > 0$ et $K_1(b) < K_2(b)$. Ces trois conditions, ainsi que la condition $0 < b < 1$, sont équivalentes à :

$$\max\left(0, \frac{\ln\left((\overline{n}(R-c) - n)/(R-1)\right)}{\ln(n)}, \frac{\ln\left(n - \underline{n}(R-c)\right)}{\ln(n)}, \frac{\ln\left(\overline{n} - \underline{n}\right)}{\ln(n)}\right) < b < 1,$$

sous réserve que les effectifs non nuls n_j vérifient :

$$1 \leqslant \underline{n} < \frac{n}{R-c} \qquad \text{et} \qquad \frac{n}{R-c} < \overline{n} \leqslant n. \tag{5.3.2}$$

Les inégalités larges de (5.3.2) sont toujours vraies. En effet, comme $n = \sum_{j=c+1}^{R} n_j$, avec $\underline{n} \leqslant n_j \leqslant \overline{n}$ pour tout j dans $\{c+1, \ldots, R\}$, nous avons $n \geqslant \underline{n}(R-c)$ et $n \leqslant \overline{n}(R-c)$. Pour avoir (5.3.2), il suffit donc d'écarter les cas « équirépartis », c'est-à-dire les données pour lesquelles :

$$\underline{n} = n_j = \frac{n}{R-c} = \overline{n},\ \forall j \in \{c+1, \ldots, R\}.$$

Au vu des observations et d'un raisonnement sur la vraisemblance similaire à celui effectué plus haut, nous montrons qu'il est raisonnable de poser une condition supplémentaire qui nous conforte dans l'observation réalisée de c valeurs nulles et $R-c$ valeurs non nulles.

Proposition 7. *Une condition suffisante pour que :*

$$\begin{aligned}&\mathbb{P}(N_1 = 0, \ldots, N_c = 0, N_{c+1} = n_{c+1}, \ldots, N_R = n_R) \geqslant \\ &\quad \mathbb{P}(N_1 = n'_1, \ldots, N_m = n'_m, N_{m+1} = 0, \ldots, N_c = 0, N_{c+1} = n'_{c+1}, \ldots, N_R = n'_R),\end{aligned}$$

pour tout entier $m \leqslant c$ et tous n'_j tels que $n'_j \leqslant n_j,\ \forall j \in \{c+1, \ldots, R\}$ et

$$\sum_{i=1}^{m} n'_i + \sum_{j=c+1}^{R} n'_j = n,$$

est la condition :

$$p_i \leqslant \frac{p_j}{n}, \quad \forall i \in \{1, \ldots, c\}, \quad \forall j \in \{c+1, \ldots, R\}. \tag{5.3.3}$$

Démonstration.
Nous supposons que m valeurs parmi les c valeurs $n_i, 1 \leqslant i \leqslant c$ sont modifiées, par exemple les premières. Soit ensuite $n'_j, j \in \{c+1, \ldots, R\}$ comme dans l'énoncé de la proposition, venant compenser ces modifications. Ainsi, nous pouvons écrire en détaillant les probabilités multinomiales que :

$$\begin{aligned}\mathbb{P}(N_1 = 0, \ldots, N_c = 0,\ N_{c+1} = n_{c+1}, \ldots, N_R = n_R) \geqslant \\ \mathbb{P}(N_1 = n'_1, \ldots, N_m = n'_m,\ N_{m+1} = 0, \ldots, N_c = 0, \\ N_{c+1} = n'_{c+1}, \ldots, N_R = n'_R)\end{aligned} \quad (5.3.4)$$

est équivalent à la condition suivante :

$$\frac{p_{c+1}^{(n_{c+1}-n'_{c+1})}}{(n_{c+1}-n'_{c+1}+1)!} \times \cdots \times \frac{p_R^{(n_R-n'_R)}}{(n_R-n'_R+1)!} \geqslant \frac{p_1^{n'_1}}{n'_1!} \ldots \frac{p_m^{n'_m}}{n'_m!}. \quad (5.3.5)$$

Supposons (5.3.3) vérifiée. Montrons qu'elle implique (5.3.5). Si nous appliquons la condition (5.3.3) à tous les éléments du membre de gauche de (5.3.5) avec des multiplicités telles que $\sum_{j=c+1}^{R} (n_j - n'_j) = \sum_{i=1}^{m} n'_i$, nous obtenons :

$$\frac{p_{c+1}^{(n_{c+1}-n'_{c+1})}}{n^{(n_{c+1}-n'_{c+1}+1)}} \times \cdots \times \frac{p_R^{(n_R-n'_R)}}{n^{(n_R-n'_R+1)}} \geqslant p_1^{n'_1} \ldots p_m^{n'_m}.$$

Ensuite, comme $(n_j - n'_j + 1)! \leqslant n^{(n_j-n'_j)}$ pour tout j dans $\{c+1, \ldots, R\}$ et $n'_i! \geqslant 1$ pour tout i dans $\{1, \ldots, m\}$, nous en déduisons (5.3.5), donc de façon équivalente l'inégalité de vraisemblance (5.3.4) énoncée plus haut qui nous conforte dans nos observations. □

Nous avons ainsi montré qu'il était raisonnable de poser la condition (5.3.3). Voyons à présent ce qu'elle implique pour a lorsqu'elle est appliquée à $\hat{p}$. La condition :

$$\hat{p}_i^{ab} \leqslant \frac{\hat{p}_j^{ab}}{n}, \quad \forall i \in \{1, \ldots, m\}, \quad \forall j \in \{c+1, \ldots, R\},$$

est équivalente à :

$$a \leqslant K_3(b),$$

avec

$$K_3(b) = \frac{\underline{n}(R-c) + n^b - n}{n^b(n(R-c)+c)}.$$

Nous prenons b tel que $K_3(b) > 0$, ce qui revient à demander $K_2(b) > 0$.

Nous résumons à présent les conditions sur a et b, et justifions le choix de la probabilité $\hat{p}^{ab}$. Soit :

$$b_{\min} = \max\left(0, \frac{\ln\left((\overline{n}(R-c)-n)/(R-1)\right)}{\ln(n)}, \frac{\ln\left(n-\underline{n}(R-c)\right)}{\ln(n)}, \frac{\ln\left(\overline{n}-\underline{n}\right)}{\ln(n)}\right).$$

Propriété 5. *La borne b_{min} est strictement inférieure à 1. De plus, si $\underline{n} = o(n)$ et $\overline{n} \sim n$ lorsque n tend vers $+\infty$, alors b_{min} et donc b convergent vers 1.*

Démonstration.
Les inégalités $\overline{n}(R-c) < nR$, $n - \underline{n}(R-c) < n$ et $\overline{n} - \underline{n} < n$ montrent que les trois composantes non nulles, dont $b_{\min}$ est le maximum, sont strictement inférieures à 1. Montrons qu'elles convergent toutes vers 1. À l'ordre 1 lorsque n tend vers $+\infty$, nous avons respectivement :

$$\begin{aligned}\frac{\ln\left((\overline{n}(R-c)-n)/(R-1)\right)}{\ln(n)} &= \frac{\ln(n) - \ln(R-1) + \ln(\overline{n}/n(R-c)-1)}{\ln(n)}, \\ &= 1 - \frac{\ln(R-1)}{\ln(n)} + \frac{\ln(R-c-1+(R-c)o(1))}{\ln(n)},\end{aligned}$$

avec $R-1$, $R-c$ et $R-c-1$ bornés,

$$\frac{\ln\left(n-\underline{n}(R-c)\right)}{\ln(n)} = 1 + \frac{\ln(1-\underline{n}/n(R-c))}{\ln(n)} = 1 + \frac{\ln(1-(R-c)o(1))}{\ln(n)},$$

avec $R-c$ borné, et

$$\frac{\ln\left(\overline{n}-\underline{n}\right)}{\ln(n)} = 1 + \frac{\ln(\overline{n}/n - \underline{n}/n)}{\ln(n)} = 1 + \frac{\ln(1+o(1))}{\ln(n)}.$$

Ces trois développements convergent vers 1, donc le maximum des trois composantes et de 0 également. □

Pour b tel que $b_{\min} < b < b_{\max} = 1$, la variable a doit vérifier :

$$\max\left(0, K_1(b)\right) < a < \min\left(1, K_2(b), K_3(b)\right).$$

Nous cherchons à définir une probabilité $\hat{p}^{ab}$ différant le plus possible de p^*, c'est pourquoi nous choisissons une valeur de b éloignée de $b_{\max}$. Pour cela, nous posons b comme une combinaison convexe de $b_{\min}$ et $b_{\max}$, avec un paramètre $h = 0.1$ déterminé empiriquement. Nous choisissons ensuite une valeur de a la plus grande possible dans l'intervalle défini par b, soit :

$$b = hb_{\max} + (1-h)b_{\min} \quad \text{et} \quad a = \min\left(1, K_2(b), K_3(b)\right) - \epsilon,$$

où ϵ est une petite constante destinée à éviter les effets de bord.

Passons maintenant aux valeurs des corrections des statistiques de Pearson et de Kullback. Pour la première, nous prenons la valeur $Q_{p^{*0}}(\hat{p}^{ab})$, que nous notons Q^{ab} et qui donne après développement :

$$\begin{aligned} Q^{ab} &= Q_{p^{*0}}(\hat{p}^{ab}), \\ &= n\sum_{r=1}^{R}\frac{\left(\hat{p}_r^{ab}\right)^2}{p_r^{*0}} - n, \\ &= n\sum_{i=1}^{c}\frac{a^2}{p_i^{*0}} + n\sum_{j=c+1}^{R}\frac{1}{p_j^{*0}}\left(\frac{n_j}{n^b} - \frac{ac+n^{1-b}-1}{R-c}\right)^2 - n, \\ &= n^{2(1-b)}Q_{p^{*0}}(p^*) - f(a,b), \end{aligned}$$

où

$$f(a,b) = n\left(1 - n^{2(1-b)} + \frac{2n^{1-b}(ac+n^{1-b}-1)}{R-c}\sum_{j=c+1}^{R}\frac{n_j}{np_j^{*0}} - a^2\sum_{i=1}^{c}\frac{1}{p_i^{*0}} - \left(\frac{ac+n^{1-b}-1}{R-c}\right)^2\sum_{j=c+1}^{R}\frac{1}{p_j^{*0}}\right).$$

De la même manière, nous choisissons la valeur $G_{p^{*0}}(\hat{p}^{ab})$, notée G^{ab}, et qui vaut :

$$\begin{aligned} G^{ab} &= G_{p^{*0}}(\hat{p}^{ab}), \\ &= 2n\sum_{r=1}^{R}\hat{p}_r^{ab}\ln\left(\frac{\hat{p}_r^{ab}}{p_r^{*0}}\right), \\ &= 2n\sum_{i=1}^{c}a\ln\left(\frac{a}{p_i^{*0}}\right) + 2n\sum_{j=c+1}^{R}\left(\frac{n_j}{n^b} - \frac{ac+n^{1-b}-1}{R-c}\right)\ln\left(\frac{1}{p_j^{*0}}\right) \\ &\quad +2n\sum_{j=c+1}^{R}\left(\frac{n_j}{n^b} - \frac{ac+n^{1-b}-1}{R-c}\right)\ln\left(\frac{n_j}{n^b} - \frac{ac+n^{1-b}-1}{R-c}\right), \\ &= n^{1-b}G_{p^{*0}}(p^*) - g(a,b), \end{aligned}$$

où

$$g(a,b) = 2n\left(\frac{ac+n^{1-b}-1}{R-c}\sum_{j=c+1}^{R}\ln\left(\frac{n_j(R-c)-n^b(ac+n^{1-b}-1)}{p_j^{*0}n^b(R-c)}\right) - a\sum_{i=1}^{c}\ln\left(\frac{a}{p_i^{*0}}\right) - n^{1-b}\sum_{j=c+1}^{R}\frac{n_j}{n}\ln\left(\frac{n_j(R-c)-n^b(ac+n^{1-b}-1)}{n_jn^{b-1}(R-c)}\right)\right).$$

Nous montrons dans le paragraphe qui suit la convergence en loi de Q^{ab} et G^{ab} vers une loi du khi-deux.

5.3.3 Convergence en loi des statistiques corrigées

Considérons le cas de s paramètres à estimer. Nous avons vu dans la section 5.1 que les statistiques de Pearson et Kullback convergent en loi sous $\mathcal{H}_0$ vers une loi du khi-deux lorsque n tend vers $+\infty$. Nous démontrons à présent le théorème suivant :

Théorème 6. *Sous les conditions de régularité (B), l'estimateur $\hat{p}^{ab}$ défini par l'expression* (5.3.1) *est tel que :*

$$\lim_{n\to+\infty} \mathcal{L}_{\mathcal{H}_0}(Q_{p^{*0}}(\hat{p}^{ab})) = \lim_{n\to+\infty} \mathcal{L}_{\mathcal{H}_0}(G_{p^{*0}}(\hat{p}^{ab})) = \chi^2_{R-s-1}.$$

Démonstration.
Pour cela, nous nous intéressons tout d'abord au remplissage des tables lorsque n tend vers l'infini. En effet, le nombre de zéros que nous avions jusqu'ici noté c est en réalité une variable aléatoire C dépendant de n, que nous notons C_n. Nous allons montrer qu'il existe presque sûrement un effectif n_0 au-delà duquel le nombre de zéros est nul.

Lemme 1. *Le nombre de zéros C_n converge vers* 0 *presque sûrement lorsque n tend vers $+\infty$.*

Démonstration.
Commençons par montrer que :

$$C_n \xrightarrow{\mathbb{P}} 0, \quad \text{ou encore} \quad \forall \epsilon > 0, \lim_{n\to+\infty} \mathbb{P}(C_n > \epsilon) = 0.$$

Pour cela, calculons pour tout n la valeur de $\mathbb{P}(C_n = c)$ pour c prenant les valeurs $\{1, \ldots, R-1\}$. La valeur R est exclue car alors le tableau serait entièrement vide, ce qui est impossible si $n \geqslant 1$. Soit $c \in \{1, \ldots, R-1\}$ et p^0 la probabilité de la loi multinomiale sous l'hypothèse nulle :

$$p^0 = (p^0_1, \ldots, p^0_c,\ p^0_{c+1}, \ldots, p^0_R).$$

Une observation a une probabilité $q^0 = p^0_1 + \cdots + p^0_c$ d'appartenir à l'une des c premières cases, et une probabilité $1 - q^0 = p^0_{c+1} + \cdots + p^0_R$ d'appartenir à l'une des $R - c$ dernières cases. La probabilité q^0 appartient à l'intervalle $]0, 1[$. Elle n'est pas nulle car il n'y a pas de zéro structural, ainsi les probabilités des catégories sont toutes non nulles. Elle est différente de la valeur 1 car $c < R$, ce qui entraîne pour la même raison que $1 - q^0$ est non nul.

Considérons la loi binomiale $\mathcal{B}(n; q^0)$. La probabilité $\mathbb{P}(C_n = c)$ d'obtenir une table avec exactement c zéros apparaissant n'importe où est égale à la probabilité que les zéros apparaissent dans les c premières cases, multipliée par le nombre de façons de choisir les c cases nulles parmi R catégories :

$$\begin{aligned}\mathbb{P}(C_n = c) &= \binom{R}{c}\mathbb{P}(N_1 = 0, \ldots, N_c = 0,\ N_{c+1} \neq 0, \ldots, N_R \neq 0), \\ &= \binom{R}{c}(1 - q^0)^n.\end{aligned}$$

Pour tout ϵ strictement positif :

$$\mathbb{P}(C_n > \epsilon) \leqslant \sum_{c=1}^{R-1} \mathbb{P}(C_n = c).$$

La probabilité $\mathbb{P}(C_n = c)$ converge vers 0 lorsque n tend vers $+\infty$ pour tout c tel que $1 \leqslant c \leqslant R-1$ parce que $\binom{R}{c}$ est borné et $(1-q^0)^n$ converge vers 0. En effet, rappelons que $q^0 \in]0, 1[$. La somme de ces probabilités converge donc également vers 0, et $\lim_{n\to+\infty} \mathbb{P}(C_n > \epsilon) = 0$.

Nous en déduisons que C_n converge vers 0 presque sûrement. Pour cela, nous utilisons le lemme de Borel-Cantelli. Posons $\epsilon > 0$. Alors :

$$\sum_{n\geqslant 1} \mathbb{P}(C_n > \epsilon) \leqslant \sum_{n\geqslant 1} \mathbb{P}(C_n \geqslant 1) = \sum_{c=1}^{R-1} \binom{R}{c} \frac{1}{q^0} < +\infty,$$

et nous concluons à la convergence presque sûre de C_n vers 0.

□

Revenons à la démonstration du théorème. Nous déduisons du lemme qu'il existe un ensemble Ω' de probabilité 1 sur lequel il existe un rang n_0 tel que pour tout $n \geqslant n_0$, $C_n = 0$. Sur Ω' et pour $n \geqslant n_0$, la variable a est posée égale à 0 et la variable b à 1, et les estimations p^* et $\hat{p}^{ab}$ sont égales, entraînant respectivement l'égalité des statistiques Q^{ab} et Q, ainsi que G^{ab} et G. Les statistiques Q et G convergeant en loi vers une loi du khi-deux à $R - s - 1$ degrés de liberté, il en est de même pour les statistiques corrigées Q^{ab} et G^{ab}.

□

5.3.4 Simulations

5.3.4.1 Quantiles et risque empirique de première espèce

Les courbes des graphiques 5.2 et 5.3 représentent les quantiles d'ordre 0.95 des statistiques Q, Q^{ab}, G, G^{ab}, G^{Ku} et $RC^{2/3}$ en fonction du nombre de zéros, sous les hypothèses nulles f_1 à f_6. Elles sont tracées pour 1 000 échantillons de $n = 400$ individus et $R = 100$ catégories, soit 99 degrés de liberté. Les effectifs des sous-échantillons permettant le calcul des quantiles figurent dans les tables 5.2 à 5.7 du paragraphe 5.2.2. La statistique G^{Ku} est représentée à titre indicatif. Nous rappelons qu'elle n'est pas définie de façon aussi rigoureuse que les autres, et qu'elle peut prendre des valeurs négatives.

Nous observons sur les courbes 5.2 et 5.3 que pour f_1, f_2 et f_3, les statistiques Q^{ab} et G^{ab} conduisent presque toujours à rejeter l'hypothèse nulle. Les variations brusques aux extrémités des courbes proviennent des petits effectifs sur lesquels ont été calculés les quantiles. Pour f_4, f_5 et f_6 en revanche, nos statistiques corrigées Q^{ab} et G^{ab} conduisent presque toujours à accepter l'hypothèse nulle, et le test est très conservateur. Nous avons corrigé la tendance à l'augmentation de Q, et nos deux statistiques sont stables. Ceci est confirmé par le calcul des risques empiriques de première espèce pour toutes les statistiques, présentés dans le tableau 5.9. De façon générale, nos corrections Q^{ab} et G^{ab} diminuent ou augmentent les statistiques Q et G, alors que la correction définie par Ku ne peut que diminuer la valeur de G.

5.3.4.2 Puissance

Nous calculons ici quelques valeurs empiriques de la puissance. Pour $i \in \{1, \ldots, 6\}$ et $j \in \{1, \ldots, 100\}$, notons $f_i(j)$ les coordonnées de f_i. Nous générons des séquences selon les lois multinomiales de probabilités f_1 à f_6, puis nous testons l'adéquation aux lois multinomiales de paramètres f_1^{bis} à f_6^{bis}, présentées dans la table 5.10.

Lorsque c est faible, la puissance est plutôt plus élevée pour les statistiques corrigées que pour les statistiques non corrigées. Ceci est la contrepartie des risques de première espèce élevés que nous avons mis en évidence dans le paragraphe précédent. Inversement, lorsque le nombre de zéros et le risque de première espèce augmentent, la puissance diminue jusqu'à devenir nulle.

Ces résultats empiriques sont prometteurs, et devront être confirmés par une étude théorique du comportement des statistiques corrigées sous $\mathcal{H}_0$ et $\mathcal{H}_0^{\mathrm{bis}}$. Les simulations et estimations que nous avons faites n'ont en effet été réalisées que pour quelques exemples de $\mathcal{H}_0$ et $\mathcal{H}_0^{\mathrm{bis}}$, et ne permettent pas de généraliser les résultats obtenus.

α	$\mathcal{H}_0$	$mode(c)$	Q	Q^{ab}	G	G^{ab}	$RC^{2/3}$
0.01	f_1	1	0	0	0	0	0
	f_2	14	0	0.411	0.005	0.717	0
	f_3	27	0.029	0.048	0	0.119	0
	f_4	36	0.051	0.005	0	0.014	0
	f_5	59	0.067	0	0	0	0.009
	f_6	79	0.034	0	0	0	0
0.05	f_1	1	0.027	0.027	0.027	0.027	0.017
	f_2	15	0.061	0.635	0.061	0.906	0.044
	f_3	26	0.086	0.129	0.009	0.243	0.033
	f_4	37	0.061	0.019	0	0.038	0
	f_5	57	0.168	0	0.011	0	0.063
	f_6	80	0.089	0	0	0	0
0.1	f_1	1	0.064	0.064	0.070	0.070	0.051
	f_2	15	0.082	0.723	0.123	0.923	0.065
	f_3	26	0.141	0.224	0.009	0.393	0.041
	f_4	37	0.059	0.059	0	0.078	0.014
	f_5	58	0.151	0	0	0	0.040
	f_6	79	0.128	0	0	0	0.014

TABLE 5.9 – Comparaison des risques empiriques de première espèce pour les statistiques Q, Q^{ab}, G, G^{ab} et $RC^{2/3}$, pour 1 000 échantillons de $n = 400$ individus répartis dans $R = 100$ catégories, aux seuils 0.01, 0.05, 0.1, pour les lois multinomiales de probabilités f_1 à f_6 et le mode du nombre de zéros correspondant.

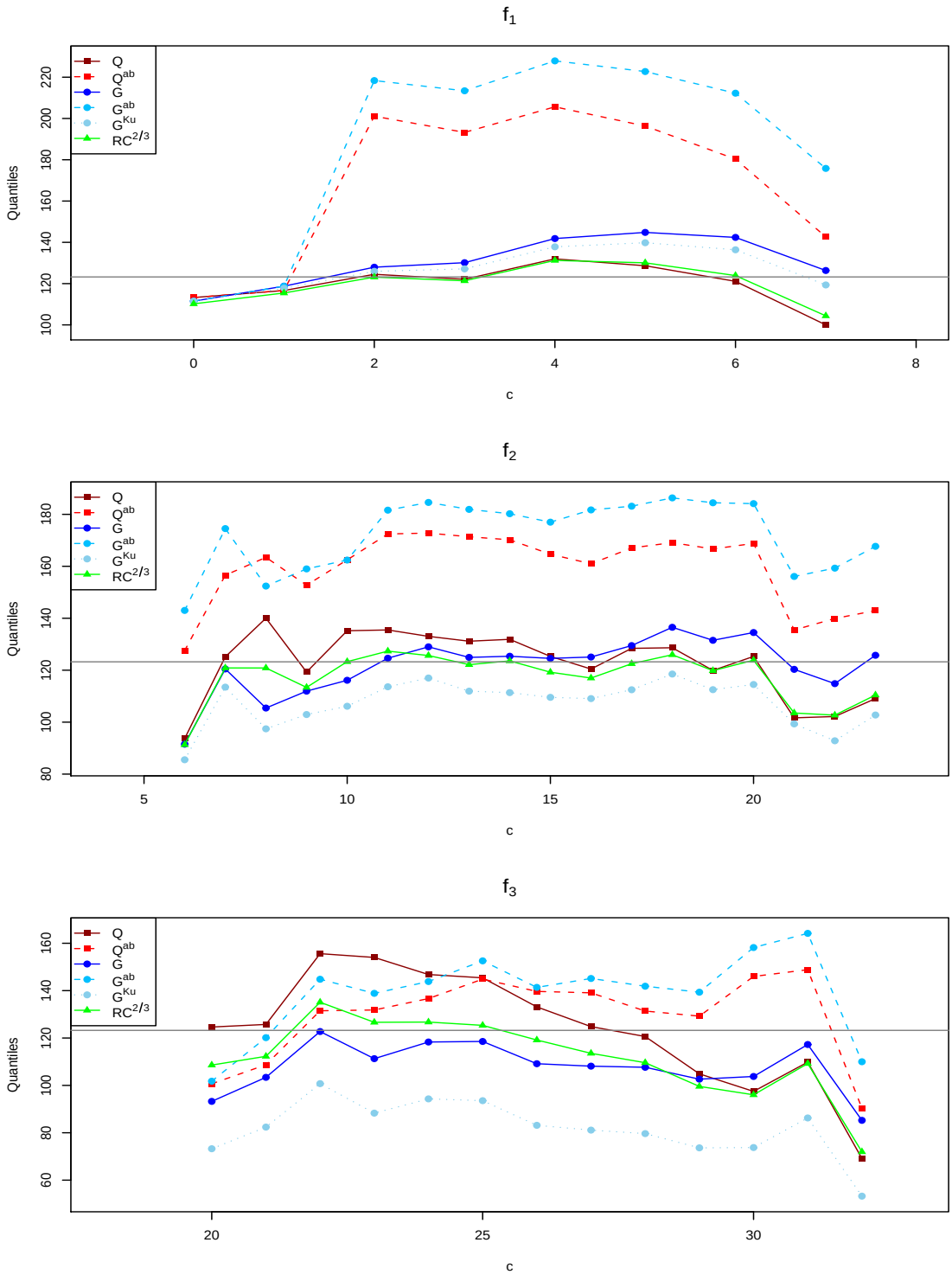


Figure 5.2 – Quantiles d'ordre 0.95 de Q, Q^{ab}, G, G^{ab}, G^{Ku} et $RC^{2/3}$ en fonction de c, sous les hypothèses nulles f_1 à f_3, pour 1 000 échantillons de 400 individus et $R = 100$ catégories. La ligne horizontale représente le seuil critique $\chi^2_{0.95,99}$.

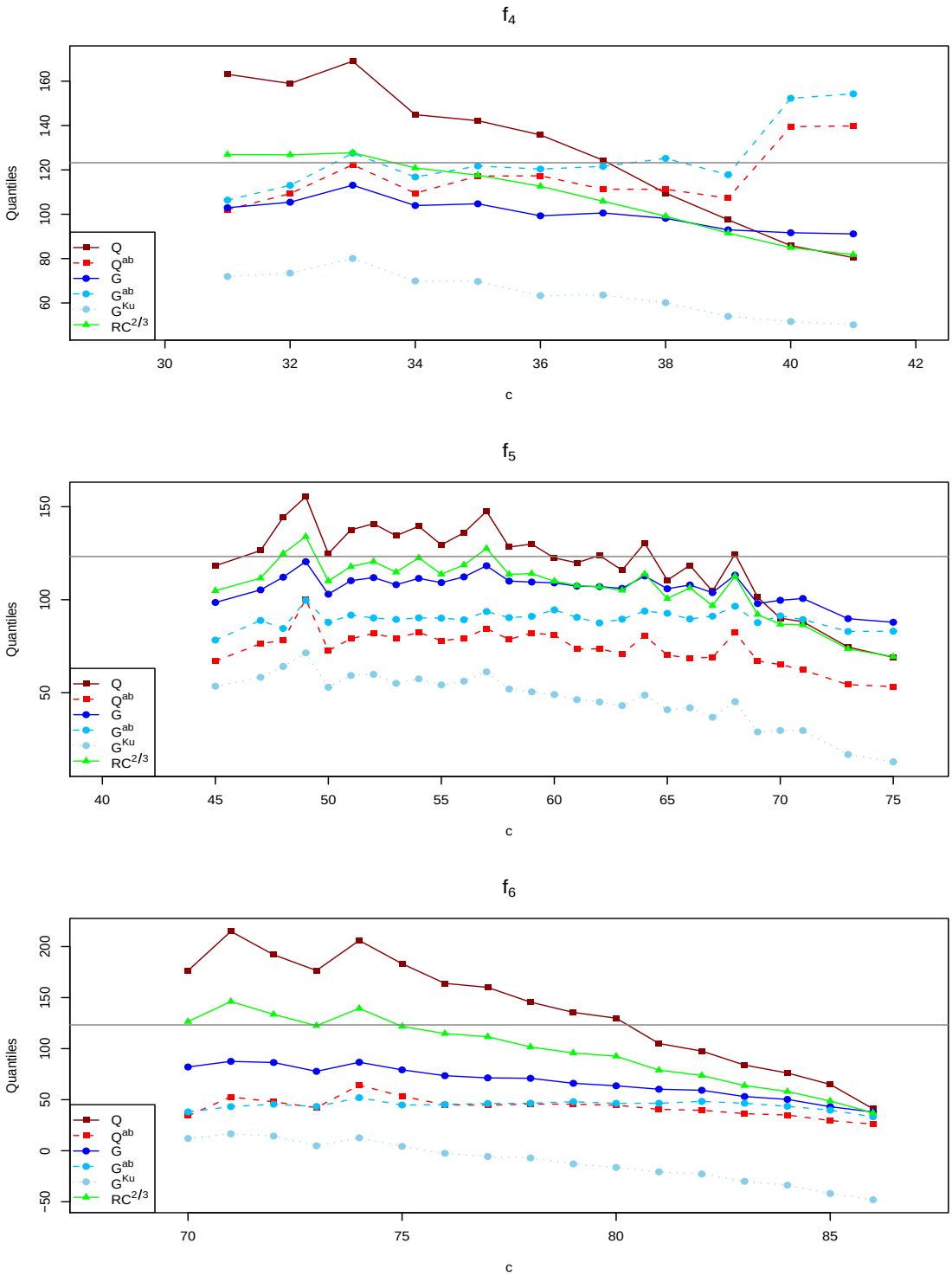


FIGURE 5.3 – Quantiles d'ordre 0.95 de Q, Q^{ab}, G, G^{ab}, G^{Ku} et $RC^{2/3}$ en fonction de c, sous les hypothèses nulles f_4 à f_6, pour 1 000 échantillons de 400 individus et $R = 100$ catégories. La ligne horizontale représente le seuil critique $\chi^2_{0.95,99}$.

$\mathcal{H}_0^{\text{bis}}$	Probabilité
f_1^{bis}	$(\underbrace{f_1(j)-0.005}_{j\in\{1,\ldots,10\}}, \underbrace{f_1(j)+0.005}_{j\in\{11,\ldots,20\}}, f_1(21),\ldots,f_1(100))$
f_2^{bis}	$(\underbrace{f_2(j)-0.0005}_{j\in\{1,\ldots,5\}}, \underbrace{f_2(j)+0.0005}_{j\in\{6,\ldots,10\}}, f_2(11),\ldots,f_1(100))$
f_3^{bis}	$(f_3(1),\ldots,f_3(80), \underbrace{f_3(j)+0.0002}_{j\in\{81,\ldots,90\}}, \underbrace{f_3(j)-0.0002}_{j\in\{91,\ldots,100\}})$
f_4^{bis}	$(f_4(1),\ldots,f_4(98), f_4(99)+0.01, f_4(100)-0.01)$
f_5^{bis}	$(\underbrace{f_5(j)-0.01}_{j\in\{1,\ldots,3\}}, \underbrace{f_5(j)+0.01}_{j\in\{4,\ldots,6\}}, f_5(7),\ldots,f_5(100))$
f_6^{bis}	$(f_6(1)-0.0001, f_6(2)+0.0001, f_6(3),\ldots,f_6(100))$

Table 5.10 – Probabilités multinomiales f_1^{bis} à f_6^{bis}.

Simulation	$\mathcal{H}_0^{\text{bis}}$	$mode(c)$	Q	Q^{ab}	G	G^{ab}	$RC^{2/3}$
f_1	f_1^{bis}	1	0.563	0.563	0.462	0.462	0.478
f_2	f_2^{bis}	15	0.096	0.659	0.068	0.892	0.039
f_3	f_3^{bis}	27	0.049	0.139	0.009	0.302	0.022
f_4	f_4^{bis}	37	0.184	0.089	0.004	0.098	0.013
f_5	f_5^{bis}	59	0.095	0	0.011	0	0.021
f_6	f_6^{bis}	79	0.094	0	0	0	0.007

Table 5.11 – Comparaison des puissances empiriques pour les statistiques Q, Q^{ab}, G, G^{ab} et $RC^{2/3}$, pour 1 000 échantillons de $n = 400$ individus répartis dans $R = 100$ catégories, au seuil 0.05 et pour le mode du nombre de zéros correspondant. Les vecteurs multinomiaux sont générés selon les probabilités f_1 à f_6 (Simulation). L'adéquation est testée pour les probabilités f_1^{bis} à f_6^{bis} ($\mathcal{H}_0^{\text{bis}}$).

Chapitre 6

Applications

Les tests issus des statistiques corrigées Q^{ab} et G^{ab} sont appliqués à différents types de données. Parmi elles, les tables de contingence en dimension trois découlant de données génomiques d'alignement de séquences non codantes, qui constituent notre motivation biologique, mais aussi deux exemples en dimension deux issus de données de nature différente.

Avant d'appliquer le test utilisant les nouvelles corrections à un jeu de données de dimension supérieure ou égale à deux, nous traitons la table en retirant les zéros y apparaissant de façon trop « homogène », au sens que nous expliquons.

L'hypothèse nulle composée des p_r^{*0}, $r \in \{1, \dots, R\}$ s'exprime à l'aide de probabilités marginales, dont nous calculons des estimations du maximum de vraisemblance à partir de comptages effectués dans la table. A l'aide d'un parallèle avec la théorie des **modèles log-linéaires** présentée dans [21], nous imposons que ces estimations soient non nulles, et retirons ainsi de la table les strates de zéros correspondant à ces marginales. Ce traitement enlève de façon structurée une partie des zéros, qui sont alors ignorés comme les zéros structuraux. La table « nettoyée » est alors considérée comme la nouvelle table d'intérêt, sur laquelle sont calculés statistiques et degrés de liberté. De façon générale certains zéros subsistent, et nous appliquons alors le test adapté au nombre de zéros restants. Par exemple, dans une table de dimension deux de taille $I \times J$ et pour le test d'indépendance :

$$\mathcal{H}_0 : \quad p_{ij} = p_{i+}p_{+j} \quad \text{contre} \quad \mathcal{H}_1 : \quad \exists\, (i_1, j_1) \in I \times J,\, p_{i_1 j_1} \neq p_{i_1 +} p_{+ j_1},$$

où les probabilités marginales sont définies de la même manière que les effectifs marginaux par sommation, nous retirons les lignes i de $\{1, \dots, I\}$ telles que $n_{i+} = 0$, et les colonnes j de $\{1, \dots, J\}$ telles que $n_{+j} = 0$.

6.1 Exemple en dimension trois : évolution de triplets de sites nucléiques

Les corrections définies précédemment sont appliquées au test de l'indépendance mutuelle de l'évolution des triplets de sites d'une séquence d'ADN. Le cadre du test est celui décrit dans la section 4.1. Nous considérons l'écriture sous la forme d'une table de taille $16 \times 16 \times 16$. Dans le cas où les effectifs sont répartis de telle façon que la table n'ait pas besoin de nettoyage, c'est-à-dire lorsque les marginales d'intérêt sont toutes non nulles, le nombre de catégories est $R = 4\ 096$. L'hypothèse d'indépendance mutuelle des caractères $\mathcal{H}_0$ et l'hypothèse alternative $\mathcal{H}_1$ s'écrivent :

$$\mathcal{H}_0 : \quad p_{ijk} = p_{i++}p_{+j+}p_{++k},$$

contre

$$\mathcal{H}_1 : \quad \exists\ (i_1, j_1, k_1) \in \{1, \ldots, 16\}^3,\ p_{i_1 j_1 k_1} \neq p_{i_1++}p_{+j_1+}p_{++k_1}.$$

Les probabilités marginales sont alors les paramètres inconnus à estimer :

$$\theta = \{p_{1++}, \ldots, p_{I++},\ p_{+1+}, \ldots, p_{+J+},\ p_{++1}, \ldots, p_{++K}\} \in [0, 1]^{64}.$$

Les trois contraintes :

$$\sum_{i=1}^{16} p_{i++} = \sum_{j=1}^{16} p_{+j+} = \sum_{k=1}^{16} p_{++k} = 1$$

impliquent un nombre $s = (16 - 1) + (16 - 1) + (16 - 1) = 3 * 16 - 3 = 45$ de paramètres inconnus. Le nombre de degrés de liberté de la loi limite du khi-deux est donc $R - s - 1 = 4\ 096 - 45 - 1 = 4\ 050$. Cependant, dans le cas de données réelles, un nettoyage est souvent nécessaire comme nous le montrons sur l'exemple qui suit.

Nous utilisons ici l'alignement des pseudogènes de la bêta-globine pour le Chimpanzé et le Gorille, présenté au paragraphe 3.3.4 du chapitre 3. La table « brute » issue de l'alignement est de taille $16 \times 16 \times 16$. Elle possède 6 strates de zéros pour le premier caractère, 6 strates pour le second et 6 strates pour le troisième. Ces couches sont retirées pendant le nettoyage. La table nettoyée de ses colonnes vides est ainsi de taille $10 \times 10 \times 10$, soit au total $R = 1\ 000$ catégories, avec $c = 898$ zéros. Dans cette table, 936 cases ont un effectif théorique E_r strictement inférieur à 0.5. Elle contient $n = 715$ individus. Les degrés de liberté sont au nombre de 972. Les valeurs des statistiques à comparer au quantile du khi-deux $\chi^2_{0.95,972} = 1\ 045.64$ sont :

$$Q = 1\ 087.01, Q^{ab} = 249.03, G = 337.31, G^{ab} = 299.09, RC^{2/3} = 498.87.$$

Seule la statistique de Pearson conduit à rejeter l'hypothèse nulle d'indépendance, et nous savons qu'en présence de zéros, elle a tendance à augmenter. En effet, comme

nous l'avons mentionné au paragraphe 5.2.2, la statistique Q converge rapidement vers sa loi asymptotique. La moyenne de la loi du khi-deux est ici de 972, ce qui peut expliquer la valeur élevée de Q. Cette valeur est beaucoup plus élevée que toutes les autres, donc nous pensons qu'elle est aberrante, et que les corrections conduisent à la bonne décision.

6.2 Exemples en dimension deux

6.2.1 Approche multi-marqueurs pour la sclérodermie systémique

Nous appliquons ici le test à une table issue d'une étude d'association cherchant à déceler chez l'Homme une association entre un gène et une maladie. Nous considérons des sites variables dans le gène, appelés **marqueurs** ou SNP pour Single Nucleotide Polymorphism.

Chaque chromosome d'une paire homologue possède pour chaque marqueur un variant, ou **allèle**. Un **haplotype** est la combinaison allélique des marqueurs d'un chromosome. Nous étudions ici trois marqueurs du gène **TNFAIP3** possédant chacun deux allèles. Neuf configurations haplotypiques sont donc possibles. La maladie potentiellement liée à la configuration du gène est la **sclérodermie systémique**. Cette maladie rare est caractérisée par une fibrose de la peau et des vaisseaux. Dans le cas du gène $TNFAIP3$, seuls huit haplotypes notés H1 à H8 sont observés.

Un **diplotype** est la combinaison des deux haplotypes d'une paire de chromosomes homologues, composée d'un chromosome maternel et d'un chromosome paternel. Les diplotypes sont donc ici au nombre de $8^2 = 64$, et notés Hi/Hj pour la combinaison de l'haplotype Hi avec l'haplotype Hj. Les marqueurs sont identifiés dans deux échantillons : un échantillon formé d'individus non affectés par cette maladie, dits **sains**, et un échantillon formé d'individus atteints par la maladie, dits **malades**.

La table diplotypique marqueur-maladie, issue de [42] et nettoyée de ses colonnes vides, est donnée en table 6.1. Nous testons l'indépendance entre la configuration diplotypique et le statut sain ou malade des individus :

$$\mathcal{H}_0 : p_{ij} = p_{i+}p_{+j} \quad \text{contre} \quad \mathcal{H}_1 : \exists\, (i_1, j_1), p_{i_1 j_1} \neq p_{i_1 +} p_{+ j_1}.$$

La table contient $n = 794$ individus. Elle est de taille 2×16, soit au total $R = 32$ catégories, avec $c = 1$ zéro. Dans cette table, seize cases ont un effectif théorique E_r strictement inférieur à 5. Les degrés de liberté sont au nombre de 15. Les valeurs des statistiques à comparer au quantile du khi-deux $\chi^2_{0.95,15} = 24.99$ sont :

$$Q = 14.62, Q^{ab} = 20.76, G = 15.82, G^{ab} = 28.43, RC^{2/3} = 14.85.$$

Statut \ Diplotype	H1/H1	H1/H2	H1/H3	H1/H4	H1/H5	H1/H6
Sain	98	7	116	2	71	3
Malade	91	9	104	3	70	12
	H2/H3	H2/H5	H2/H6	H3/H3	H3/H4	H3/H5
Sain	4	2	0	34	1	42
Malade	5	4	1	30	2	40
	H3/H6	H4/H5	H5/H5	H5/H6		
Sain	2	1	13	1		
Malade	7	1	13	5		

Table 6.1 – Table diplotypique pour l'étude de l'association entre trois marqueurs du gène $TNFAIP3$ et la sclérodermie systémique.

Seule la statistique G^{ab} conduit à rejeter l'hypothèse d'indépendance des deux caractères. Des études préalables exposées dans [42] ont été menées sur chaque marqueur séparément, ou sur la table haplotypique de dimension 2×8. Elles ont montré une association significative du gène et de la maladie, ce que confirme la statistique G^{ab} de notre analyse diplotypique.

En pratique, les tables diplotypiques sont difficiles à manipuler en raison du nombre important de zéros. Elles sont pourtant intéressantes parce qu'elles tiennent compte à la fois de l'information apportée par la combinaison de plusieurs marqueurs, et de l'information combinée des deux chromosomes homologues. Les statistiques corrigées que nous proposons peuvent dans ce cas s'avérer utiles pour analyser les tables diplotypiques, particulièrement creuses, qui sont actuellement difficiles à analyser avec les statistiques de test classiques.

6.2.2 Trophie et végétaux des cours d'eau de la Petite Camargue Alsacienne

Nous étudions ici deux caractères liés à la végétation dans les cours d'eau : la trophie et la composition en espèces végétales. Trois catégories de **trophie** sont distinguées selon la richesse du milieu en éléments minéraux nutritifs. Un milieu est dit **oligotrophique**, **mésotrophique** ou **eutrophique** selon qu'il est pauvre, moyennement riche ou riche en substances nutritives.

Les espèces végétales observées dans les cours d'eau de la Petite Camargue Alsacienne et qui ne sont pas communes sont **rares**, **exotiques** ou **polluo-tolérantes**. Une espèce est rare lorsqu'elle est très peu fréquente en Alsace, ou lorsqu'elle est inscrite sur la liste rouge d'Alsace. Elle est exotique lorsqu'elle est originaire d'une

Trophie \ (r,p,e)	$(0,0,0)$	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$	$(1,1,0)$	$(0,1,1)$
Oligotrophique	0	0	3	0	3	2
Mésotrophique	2	1	0	2	1	0
Eutrophique	2	0	3	1	1	0

TABLE 6.2 – Table de contingence pour l'étude conjointe de la trophie et de la composition en espèces végétales.

autre région. Enfin, elle est polluo-tolérante si elle est connue pour résister à des niveaux élevés de pollution.

Nous attribuons à chaque cours d'eau trois caractéristiques notées (r,p,e), indiquant respectivement la présence (1) ou l'absence (0) d'espèces de type rare, exotique ou polluo-tolérant, destinées à caractériser sa composition végétale. Les données ont été recueillies dans 21 cours d'eau différents de la Petite Camargue Alsacienne et à des dates différentes, et sont issues de [76].

Nous testons l'indépendance entre le niveau trophique et la composition végétale :

$$\mathcal{H}_0 : p_{ij} = p_{i+}p_{+j} \quad \text{contre} \quad \mathcal{H}_1 : \exists\, (i_1,j_1), p_{i_1j_1} \neq p_{i_1+}p_{+j_1}.$$

La table « brute » possède 2 colonnes de zéros, qui sont retirées pendant le nettoyage. La table nettoyée de ses colonnes vides est donnée en table 6.2.
Elle contient $n = 21$ individus. Elle est de taille 3×6, soit au total $R = 18$ catégories, avec $c = 7$ zéros. Dans cette table, trois cases ont un effectif théorique E_r strictement inférieur à 0.5. Les degrés de liberté sont au nombre de 10. Les valeurs des statistiques à comparer au quantile du khi-deux $\chi^2_{0.95,10} = 18.31$ sont :

$$Q = 14.38, Q^{ab} = 20.68, G = 18.67, G^{ab} = 26.05, RC^{2/3} = 14.84.$$

Les statistiques Q^{ab}, G et G^{ab} conduisent à rejeter l'hypothèse nulle d'indépendance. Il semblerait donc y avoir un lien entre trophie et composition en espèces végétales.

Plus précisément, la table nous montre que les espèces rares sont plutôt présentes dans les milieux oligotrophes. Ce phénomène est en partie lié à la nature de ces végétaux. En effet, les espèces rares ne sont souvent adaptées qu'aux milieux oligotrophes dont l'étendue diminue en Alsace depuis quelques années. D'autre part, nous remarquons que les espèces polluo-tolérantes sont majoritaires en milieu eutrophique. Ce type de milieu est soumis à une forte compétition, et ces espèces possèdent en général une bonne résistance aux conditions environnementales difficiles et concurrentielles.

6.2.3 Perspectives

Ces résultats nous encouragent donc dans l'utilisation des statistiques corrigées Q^{ab} et G^{ab} à la place de G, mais surtout de Q qui prend souvent des valeurs très élevées lorsque R et n augmentent. Outre les tests d'indépendance mutuelle en dimensions deux et trois présentés ici, ces corrections peuvent être appliquées à des tests d'indépendance partielle, marginale ou conditionnelle en dimension quelconque.

Bibliographie

[1] J. Adachi and M. Hasegawa. MOLPHY version 2.3. programs for molecular phylogenetics based on maximum likelihood. In M. Ishiguro, G. Kitagawa, Y. Ogata, H. Takagi, Y. Tamura, and T. Tsuchiya, editors, *Computer Science Monographs*, number 28. The Institute of Statistical Mathematics, Tokyo, 1996.

[2] A. Agresti. *Categorical data analysis*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1990.

[3] A. Agresti. *An introduction to categorical data analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, second edition, 2007.

[4] A. Albert. Estimating the infinitesimal generator of a continuous time, finite state Markov process. *Ann. Math. Statist.*, 33 :727–753, 1962.

[5] F. Antequera and A. Bird. CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.*, 9 :661–667, 1999.

[6] P. F. Arndt, C. B. Burge, and T. Hwa. DNA sequence evolution with neighbor-dependent mutation. *J. Comput. Biol.*, 10(3/4) :313–322, 2003.

[7] P. F. Arndt and T. Hwa. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10) :2322–2328, 2005.

[8] J. M. Bahi and C. J. Michel. A stochastic model of gene evolution with time dependent pseudochaotic mutations. *Bull. Math. Biol.*, 71(3) :681–700, 2009.

[9] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37 :1554–1563, 1966.

[10] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41 :164–171, 1970.

[11] E. Benard and C. J. Michel. Research Article : Computation of direct and inverse mutations with the SEGM web server (Stochastic Evolution of Genetic Motifs) : An application to splice sites of human genome introns. *Comput. Biol. Chem.*, 33(4) :245–252, 2009.

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

[13] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3) :179–195, 1975.

[14] M. W. Birch. A new proof of the Pearson-Fisher theorem. *Ann. Math. Statist*, 35 :817–824, 1964.

[15] J. Bérard, J.-B. Gouéré, and D. Piau. Solvable models of neighbor-dependent nucleotide substitution processes. *Mathematical Biosciences*, 211 :56–88, 2008.

[16] M. Bulmer. Neighboring base effects on substitution rates in pseudogenes. *Mol. Biol. Evol.*, 3(4) :322–329, 1986.

[17] D. Charif and J.R. Lobry. SeqinR 1.0-2 : a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In H.E. Roman U. Bastolla, M. Porto and M. Vendruscolo, editors, *Structural approaches to sequence evolution : Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, 2007.

[18] F.-C. Chen and W.-H. Li. Genomic divergences between Humans and other hominoids and the effective population size of the common ancestor of Humans and Chimpanzees. *Am. J. Hum. Genet.*, 68 :444–456, 2001.

[19] O. F. Christensen. Pseudo-likelihood for non-reversible nucleotide substitution models with neighbour dependent rates. *Stat. Appl. Genet. Mol. Biol.*, 5(1) :Article 18, 2006.

[20] O. F. Christensen, A. Hobolth, and J. L. Jensen. Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *J. Comput. Biol.*, 12(9) :1166–1182, 2005.

[21] R. Christensen. *Log-linear models and logistic regression*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1997.

[22] G. A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, 51(1) :79–94, 1989.

[23] W. G. Cochran. The χ^2 test of goodness of fit. *Ann. Math. Statistics*, 23 :315–345, 1952.

[24] W. J. Conover. *Practical nonparametric statistics*. John Wiley & Sons, 1999.

[25] Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[26] N. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B*, 46(3) :440–464, 1984.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1) :1–38, 1977.

[28] J. L. Doob. *Stochastic processes*. John Wiley & Sons Inc., New York, 1953.

[29] L. Duret and N. Galtier. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Mol. Biol. Evol.*, 17(11) :1620–1625, 2000.

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

[31] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, 2003.

[32] A. Finkler. Goodness of fit statistics for sparse contingency tables. *Submitted*.

[33] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *J. Roy. Stat. Soc.*, 85(1) :87–94, 1922.

[34] W. Fitch and E. Markowitz. 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 :579–593, 1970.

[35] N. Galtier. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18(5) :866–873, 2001.

[36] G. Gibson and S. V. Muse. *Précis de génomique*. De Boeck, 2004.

[37] E. J. Gilbert. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.*, 30 :688–697, 1959.

[38] N. Goldman. Statistical tests of models of DNA substitution. *J. Mol. Evol.*, 36 :182–198, 1993.

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

[40] I. J. Good. *Probability and the weighing of evidence*. Charles Griffin & Co. Ltd., London, 1950.

[41] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154) :862–864, 1974.

[42] M. Guedj et al. Association of TNFAIP3 rs5029939 variant with systemic sclerosis in European Caucasian population. *Under review*.

[43] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5) :696–704, 2003.

[44] M. Hasegawa, H. Kishino, and T. Yano. Dating the Human-Ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22 :160–174, 1985.

[45] S. T. Hess, J. D. Blake, and R. D. Blake. Wide variations in neighbor-dependent substitution rates. *J. Mol. Evol.*, 236 :1022–1033, 1994.

[46] J. L. Jensen and A.-M. K. Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. in Appl. Probab.*, 32(2) :499–517, 2000.

[47] T. H. Jukes and C. R. Cantor. *Evolution of protein molecules.* Academy Press, 1969.

[48] F. P. Kelly. *Reversibility and stochastic networks.* John Wiley & Sons Ltd., Chichester, 1979.

[49] S.-H. Kim, H. Choi, and S. Lee. Estimate-based goodness-of-fit test for large sparse multinomial distributions. 53 :1122–1131, 2009.

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

[51] K. J. Koehler and K. Larntz. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Am. Stat. Assoc.*, 75 :336–344, 1980.

[52] H. H. Ku. A note on contingency tables involving zero frequencies and the 2Î test. *Technometrics*, 5(3) :398–400, 1963.

[53] S. Kullback. *Information theory and statistics.* John Wiley and Sons, Inc., New York, 1959.

[54] S. Kumar and S. B. Hedges. A molecular timescale for molecular evolution. *Nature*, 392 :917–920, 1998.

[55] F. Larsen, G. Gundersen, R. Lopez, and H. Prydz. CpG islands as gene markers in the human genome. *Genomics*, 13(4) :1095–1107, 1992.

[56] S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series.* The Clarendon Press Oxford University Press, New York, 1996.

[57] G. Lunter and J. Hein. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20(1) :216–223, 2004.

[58] I. Miklos, G. A. Lunter, and I. Holmes. A "long indel" model for evolutionary sequence alignment. *Mol. Biol. Evol.*, 21(3) :529–540, 2004.

[59] B. R. Morton. The influence of neighboring base composition in substitution in plant chloroplast coding sequences. *Mol. Biol. Evol.*, 14(2) :189–194, 1997.

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

[61] J. Neyman and E. S. Pearson. Further notes on the χ^2 distribution. *Biometrika*, 22 :298–305, 1931.

[62] P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S. D. Ehrlich, B. Prum, and P. Bessières. Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic Acids Research*, 30(6) :1418–1426, 2002.

[63] H. Ochman and A. C. Wilson. Evolution in bacteria : evidence for a universal substitution rate in cellular genomes. *J Mol Evol*, 26(1-2) :74–86, 1987.

[64] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50 :157–175, 1900.

[65] A.-M. K. Pedersen and J. L. Jensen. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.*, 18(5) :763–776, 2001.

[66] A.-M. K. Pedersen, C. Wiuf, and F. B. Christiansen. A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.*, 15(8) :1069–1081, 1998.

[67] T. Petrie. Probabilistic functions of finite state Markov chains. *Ann. Math. Statist*, 40 :97–115, 1969.

[68] R Development Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009.

[69] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 :257–286, 1989.

[70] T. R. C. Read and N. A. C. Cressie. *Goodness-of-fit statistics for discrete multivariate data*. Springer-Verlag, New York, 1988.

[71] L. Sachs. *Applied statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1982.

[72] A. C. Siepel and D. Haussler. Combining phylogenetic and hidden markov models in biosequence analysis. *J. Comput. Biol.*, 11(2/3) :413–428, 2004.

[73] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17, 1986.

[74] J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33 :114–124, 1991.

[75] J. L. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality : an improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34 :3–16, 1992.

[76] M. Trémolières, I. Combroux, A. Hermann, and P. Nobelis. Conservation status assessment of aquatic habitats within the Rhine floodplain using an index based on macrophytes. *Ann. Limnol.-Int. J. Lim.*, 43(4) :233–244, 2007.

[77] C. Tuffley and M. Steel. Modelling the covarion hypothesis of nucleotide substitution. *Math. Biosci*, 147 :63–91, 1998.

[78] J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507) :1304–1351, 2001.

[79] S. Whelan and N. Goldman. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.*, 16(9) :1292–1299, 1999.

[80] S. Whelan and N. Goldman. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, 167(4) :2027–2043, 2004.

[81] C.-F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1) :95–103, 1983.

[82] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol*, 10(6) :1396–1401, 1993.

[83] Z. Yang. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39 :39–105, 1994.

[84] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites : approximate methods. *J. Mol. Evol*, 39 :39–306, 1994.

[85] Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139(2) :993–1005, 1995.

[86] Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1) :431–449, 2000.

[87] Z. Yang and A. D. Yoder. Estimation of the transition/transversion rate bias and species sampling. *J. Mol. Evol.*, 48(3) :274–283, 1999.

[88] J. K. Yarnold. The minimum expectation in X^2 goodness of fit tests and the accuracy of approximations for the null distribution. *J. Amer. Statist. Assoc.*, 65 :864–886, 1970.

Dans ce travail nous étudions sous deux aspects la dépendance au contexte pour l'évolution par substitution des séquences nucléotidiques.
Dans une première partie nous définissons un modèle évolutif simple intégrant la distinction entre transitions et transversions d'une part, et une dépendance des nucléotides à leur voisin de gauche modélisant l'effet CpG d'autre part. Nous montrons que ce modèle peut s'écrire sous la forme d'une chaîne de Markov cachée et estimons ses paramètres par la mise en œuvre de l'algorithme de Baum-Welch. Nous appliquons enfin le modèle à l'estimation de taux de substitution observés dans l'évolution de séquences génétiques.
Dans une deuxième partie nous développons des corrections pour les statistiques classiques du test d'adéquation d'un échantillon à une loi multinomiale en présence de zéros aléatoires. En effet, les tests d'indépendance de l'évolution de triplets de nucléotides voisins impliquent des tables de contingence possédant de nombreuses cases nulles et se ramènent à des tests d'adéquation sur des vecteurs creux. Les statistiques de Pearson et de Kullback ne peuvent alors être employées. À partir de celles-ci, nous considérons des statistiques corrigées qui conservent le même comportement asymptotique. Nous les utilisons pour réaliser des tests d'indépendance, non seulement dans le cadre des données génomiques de la première partie, mais également pour des données écologiques et épidémiologiques.

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE
UMR 7501
Université de Strasbourg et CNRS
7 Rue René Descartes
67 084 STRASBOURG CEDEX

Tél. 03 68 85 01 29
Fax 03 68 85 03 28

www-irma.u-strasbg.fr
irma@math.unistra.fr

IRMA 2010/05
ISSN 0755-3390
http ://tel.archives-ouvertes.fr/tel-00490844