Université de Montréal

Évolution Moléculaire
Un modèle Markov-modulé pour les processus de substitution

Présenté par
Eric Fournier

Département de Biochimie et médecine moléculaire
Faculté de médecine

Mémoire présenté à la Faculté de médecine en vue de
l'obtention du grade de maîtrise ès sciences en bioinformatique

janvier, 2014

Université de Montréal

Ce mémoire intitulé

Évolution Moléculaire
Un modèle Markov-modulé pour les processus de substitution

Présenté par
Eric Fournier

a été évalué par un jury composé des personnes suivantes

Sylvie Hamel,	Président
Nicolas Lartillot,	Directeur
Hervé Philippe,	Codirecteur
B. Franz Lang,	Membre

Mémoire accepté le : ..... février, 2014 .....

RÉSUMÉ

Les processus Markoviens continus en temps sont largement utilisés pour tenter d'expliquer l'évolution des séquences protéiques et nucléotidiques le long des phylogénies. Des modèles probabilistes reposant sur de telles hypothèses sont conçus pour satisfaire la non-homogénéité spatiale des contraintes fonctionnelles et environnementales agissant sur celles-ci. Récemment, des modèles Markov-modulés ont été introduits pour décrire les changements temporels dans les taux d'évolution site-spécifiques (hétérotachie). Des études ont d'autre part démontré que non seulement la force mais également la nature de la contrainte sélective agissant sur un site peut varier à travers le temps. Ici nous proposons de prendre en charge cette réalité évolutive avec un modèle Markov-modulé pour les protéines sous lequel les sites sont autorisés à modifier leurs préférences en acides aminés au cours du temps. L'estimation *a posteriori* des différents paramètres modulants du noyau stochastique avec les méthodes de Monte Carlo est un défi de taille que nous avons su relever partiellement grâce à la programmation parallèle. Des réglages computationnels sont par ailleurs envisagés pour accélérer la convergence vers l'optimum global de ce paysage multidimensionnel relativement complexe. Qualitativement, notre modèle semble être capable de saisir des signaux d'hétérogénéité temporelle à partir d'un jeu de données dont l'histoire évolutive est reconnue pour être riche en changements de régimes substitutionnels. Des tests de performance suggèrent de plus qu'il serait mieux ajusté aux données qu'un modèle équivalent homogène en temps. Néanmoins, les histoires substitutionnelles tirées de la distribution postérieure sont bruitées et restent difficilement interprétables du point de vue biologique.

Mots clés : évolution moléculaire, inférence Bayésienne, processus de substitution, modèle Markov-modulé.

ABSTRACT

Time-continuous Markovian process are widely used to understand the mechanism of nucleotidic acids and proteins evolution along phylogeny. Already existing probabilistic models based on such hypothesis are designed to satisfy the non-homogeneity of functional and environmental constraints acting across those biological sequences. Recently, Markov-modulated models have been introduced to describe site-specific temporal rate variation (heterotachy). Moreover, studies have demonstrated that not only strength but also the nature of the constraint acting on a specific site can vary over time. Here we propose to accommodate this evolutionary reality with a Markov-modulated model for proteins under which sites are authorized to change their amino acids propensities across time. Posterior estimation of the stochastic kernel hidden parameters with Monte Carlo methods is a challenging approach that we partially overcome with parallel computing. Fine-tuning are otherwise planned to accelerate convergence toward the target posterior stationnary distribution. Qualitatively, our model seems to be able to capture temporal heterogeneity from real sequences data sets whose evolutionary history is assumed to be rich in substitutional switch events. Furthermore, evaluation of the model performance suggest that he provides a better fit to the data set than the time-homogeneous equivalent model. Nonetheless, substitutional histories sampled from the posterior distribution are quite noisy and remain difficult to interpret biologically.

Key words : molecular evolution, Bayesian inference, substitution process, Markov-modulated model.

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

AIC Akaike Information Criteria
ARN Acide RiboNucléique

BF Bayes Factor
BIC Bayesian Information Criteria
BP Break Point

CAT CATegories
cov covarion
CPO Conditionnal Predictive Ordinate

D-Gam Discrete Gamma distribution for variable substitution rates across sites

EM Expectation Maximisation
Exp Distribution Exponentielle

F81 Felsenstein 1981

GCC GNU Compiler Collection
GTR General Time Reversible
GWAS Genome-Wide Association Study

HKY85 Hasegawa, Kishino and Yano 1985
HSp3 3 Hidden States
HSp6 6 Hidden States
HSSP homology-derived secondary structure of proteins

i.i.d independant and identical distributed
IC Intervalle de crédibilité

JC69 Jukes and Cantor 1969
JTT Jones, Taylor and Thornton

K80 Kimura 1980

LBA Long Branch Attraction
LG Le and Gascuel
LOOCV Leave-One-Out Cross Validation
LPML Log PseudoMargianal Likelihood
LRT Likelihood Ratio Test

LS Least Square

MCMC Markov Chain Monte Carlo
MH Metropolis and Hastings
ML Maximum Likelihood
MM Modèle Markov-modulé
MPI Message Parsing Interface
mtREV mitochondrial REVersible

NJ Neighbor Joining
NJ Maximum Parcimony

PAM Point Accepted Mutation matrix
PIP Probability of Identical Profile

RAS Rate Across Sites
RQCHP Réseau Québécois de Calcul de Haute Performance

s.e. sous-échantillonnage

UdeM Université de Montréal
UPGMA Unweighted Pair Group Method with Arithmetic Mean

WAG Whelan And Goldman

Liste des symboles

S Nombre d'acides aminés

$\boldsymbol{Q}$ Matrice de taux de transition instantané

$\wp_i$ Proportion d'états observés dans le groupe des microsporidies qui sont dans un état caché différent de celui des autres taxons

π Fréquences d'équilibre des acides aminés

$\mathcal{F}$ Moyenne des $\wp_i$ à travers tous les sites

ρ Taux relatif déchange entre acides aminés

δ_{CPO_i} Gain moyen en CPO par site

θ Vecteur de paramètres

$\mathfrak{D}$ Densité moyenne de de modulations par site modulant

α Paramètre de forme de la distribution gamma

$\mathcal{D}$ Densité de sites modulant

κ Taux de transition versus transversion

d Densité de sites avec substitutions

μ Taux de substitution

$\mathfrak{d}$ Densité moyenne de substitutions par état caché

Γ gamma

$\mathbf{l}$ Longueur de l'arbre

ζ Stationnaire de la matrice Q

η Fréquences déquilibre des états cachés

$\mathcal{O}$ Complexité algorithmique

$\mathbf{L}$ Likelihood

ϵ Diversité nucléotidique moyenne par site

Ξ Histoire substitutionnelle

$\eth$ Taux relatif déchange entre états cachés

G Nombre d'états cachés

w_m poids de la mth catégorie de la distribution gamma discrétisée

r_m vitesse d'évolution médiane de la mth catégorie de la distribution gamma discrétisée

C_i Données au site i

l Longueur de branche

I Matrice identité

D Données

P_j Proportion de transitions entre états cachés sur la branche j

H_j Nombre de transitions entre états cachés sur la branche j

O_j Nombre de transitions entre états observés sur la branche j

REMERCIEMENTS

J'aimerais remercier premièrement mon directeur de recherche Nicolas Lartillot pour avoir accepté de me prendre sous son aile. L'expérience et les connaissances qu'il m'a permis d'acquérir durant les deux dernières années sont inestimables. Je remercie aussi mon codirecteur Hervé Philippe pour le temps qu'il m'a consacré.

Je suis également reconnaissant envers les gens du Laboratoire de santé publique du Québec sans qui je n'aurais pu plonger dans cette aventure. Un merci spécial à Mme Andrée Gilbert qui a su entamer les démarches nécessaires.

Merci à ma mère pour ses encouragements. Mais merci surtout à ma bien-aimée, Julie Forget pour son support et sa patience. Sa présence reconfortante m'a permis de passer plus aisément à travers certaines périodes plus difficiles.

1 INTRODUCTION

1.1 La Phylogénie Moléculaire : mise en contexte

En plus d'établir les relations évolutives existantes entre les différents groupes d'organismes vivants, la phylogénie moléculaire peut contribuer à l'avancement d'études pangénomiques (GWAS), permettre de reconstruire des séquences ancestrales, être utilisée pour identifier des résidus ciblés par la sélection naturelle et estimer des temps de divergence entre espèces. L'abondance des séquences génomiques et protéiques aujourd'hui disponibles dans les différentes bases de données biologiques fournit potentiellement une considérable quantité de signaux évolutifs pour les phylogénéticiens modernes. La récupération de ces signaux exige premièrement la mise au point d'algorithmes d'alignement efficaces et précis permettant de comparer les résidus homologues. Elle requiert deuxièmement l'élaboration de modèles d'évolution décrivant de façon adéquate les processus évolutifs réels. Durant les dernières décennies, plusieurs modèles ont été proposés. Leur fiabilité dépend largement de la distance évolutive entre les taxons étudiés. Un jeu de taxons, dans lequel se retrouvent des espèces relativement éloignées, requiert habituellement des modèles plus complexes prenant en compte la dynamique évolutive élaborée sous-jacente.

Les mécanismes d'évolution proposés par les premiers auteurs de la génomique évolutive étaient relativement simples. Leurs approches assumaient généralement une homogénéité temporelle (le long d'un arbre phylogénétique) et spatiale (à travers les positions, ou sites, d'un alignement de séquences génomiques ou protéiques) du processus évolutif. De telles suppositions sont cependant dans plusieurs cas incompatibles avec les données empiriques et peuvent mener à des artéfacts de reconstruction phylogénétique. Par conséquent, les modèles d'évolution mis au point aujourd'hui tentent d'assouplir de telles contraintes par l'ajout de paramètres dit d'hétérogénéité. Celui que nous proposons dans le présent projet de maîtrise ajoute un dégré d'hétérogénéité supplémentaire en autorisant des modulations de type qualitatives entre différents processus de substitution site-spécifiques au cours du temps.

Dans la suite, nous ferons d'abord un bref survol des différents modèles de reconstruction phylogénétiques proposés. Nous y introduirons les modèles probabilistes inspirés de concepts mathématiques et statistiques fondamentaux. Ceux-ci sont beaucoup plus robustes comparativement aux méthodes de distances (méthodes de moindres carrés (Cavalli-Sforza & Edwards 1967 ; Fitch & Margoliash 1967), UPGMA (Sokal & Michener 1958) et *Neighbor Joining* (Saitou & Nei 1987)) et de maximum de parcimonie (Farris 1970 ; Fitch 1971b). Ensuite, nous décrirons en détail notre contribution. L'on constatera que notre approche a fait appel à des connaissances émergeants de divers domaines scientifiques ; informatique, statistique, mathématique et biologie. La touche informatique s'est d'ailleurs révélée plus qu'essentielle pour supporter efficacement les ressources requises par l'approche Bayésienne implémentée.

Ce projet est une preuve concrète que tenter de percer le mystère des processus évolutifs derrière la merveilleuse diversité du vivant qui nous entoure exige d'avoir le courage d'ex-

plorer des horizons qui sont souvent hors de notre zone de confort.

1.2 Modélisation par Maximum de Vraisemblance versus par Inférence Bayésienne

Les séquences génétiques et protéiques observées aujourd'hui sont le résultat d'un processus évolutif stochastique complexe. De plus, le nombre fini et restreint de nucléotides (4) et d'acides aminés (20) les constituant vient compliquer sérieusement la modélisation de ce processus ; une série de substitutions multiples le long d'une branche d'arbre peut rapidement devenir saturée. Ainsi, les résidus obervés à une position précise d'un alignement, même s'ils sont identiques, n'ont pas nécessairement la même histoire évolutive dû au phénomène d'homoplasie (figure 1). Lorsque l'on cherche à établir des liens évolutifs entre espèces, dont certaines évoluent rapidement comparativement aux autres, ce phénomène est d'ailleurs à l'origine d'artéfacts de reconstruction phylogénétiques causés par l'attraction des longues branches (LBA) (Huelsenbeck 1995 ; Huelsenbeck & Hillis 1993). L'artéfact LBA étant une erreur sytématique observée lorsque deux espèces évoluant rapidement sont incorrectement placées sous un même ancêtre rapproché.

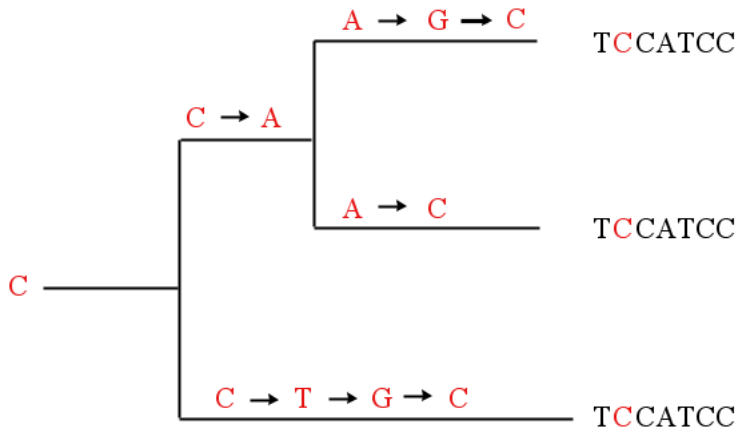


Figure 1 – Exemple d'un phénomène homoplasique moléculaire. Convergence vers la cytosine (C).

Les méthodes probabilistes peuvent définitivement mieux accommoder la réalité des processus évolutifs (Holder & Lewis 2003). Deux approches probabilistes sont utilisées en phylogénie moléculaire. Celle du maximum de vraisemblance (ML) (Felsenstein 1981) et plus récemment, l'inférence Bayésienne (Mau 1996 ; Li 1996 ; Rannala & Yang 1996 ; Yang & Rannala 1997 ; Huelsenbeck et al. 2000 ; Larget & Simon 1999). Les deux approches s'entendent sur le fait qu'il est possible d'expliquer la stochasticité du processus substitutionnel avec un modèle d'évolution paramétré ; topologie d'arbre, longueurs de branches,

taux relatifs d'échange entre nucléotides ou acides aminés, vitesse d'évolution et autres paramètres.

L'aspect sur lequel les fréquentistes et les Bayésiens ne s'entendent pas concerne l'approche utilisée pour estimer les valeurs de paramètres sachant les données empiriques. Étant donné un paramètre précis du modèle d'évolution, l'approche ML consiste essentiellement à chercher la valeur de ce paramètre qui maximise la probabilité des données observées. Ou autrement dit, celle qui a la vraisemblance maximale. Des algorithmes de *hill-climbing* (Bryant et al. 2005) sont conçus à cette fin. L'approche Bayésienne offre quant à elle une mesure plus naturelle d'incertitude associée à la valeur du paramètre sous forme d'une distribution *a posteriori*. Cela permet d'ailleurs aux modèles Bayésiens d'être plus riches en paramètres phylogénétiques et donc plus complexes. Cependant, puisque ML permet de calculer des vraisemblances pour des topologies précises ou des valeurs ponctuelles de paramètres, il est possible avec cette méthode de faire des tests de monophylie et de comparer la performance de différents modèles de façon plus simple.

Une autre différence entre les deux approches, et non la moindre, est que la méthode Bayésienne repose explicitement sur une prior. D'après le théorème de Bayes suivant, la probabilité postérieure d'un ensemble de paramètres définissant un modèle M (collectivement désignés par θ) conditionnelle aux données D est proportionnelle au produit leur vraisemblance par leur prior ;

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)p(\theta \mid M)}{p(D \mid M)}. \tag{1}$$

où

$$p(D \mid M) = \int_{\theta} p(D \mid \theta, M)p(\theta \mid M)d\theta \tag{2}$$

correspond à la probabilité marginale des données. L'incorporation d'une distribution *a priori* peut être avantageuse. Mais peut aussi au contraire mener à des estimations postérieures incorrectes. L'indécision implique souvent de s'en remettre à une prior uniforme non-informative (Berger 1985).

L'échantillonnage de la distribution postérieure avec les chaînes de Markov Monte Carlo (MCMC) appliquées à la phylogénie moléculaire (Rannala 2002 ; Larget & Simon 1999 ; Mau 1996 ; Li 1996) est ce qui permet d'élaborer des modèles Bayésiens de dimensions supérieures sans qu'ils soient sur-ajustés aux données empiriques (Felsenstein 2004). Les deux méthodes MCMC les plus fréquemment utilisées sont l'algorithme de Metropolis-Hastings (MH) (Metropolis et al. 1953 ; Hastings 1970) et le *Gibbs sampling* (Geman & Geman 1984). Avec MH, les mouvements proposés sur θ (θ^*) sont acceptés avec une probabilité Min(1, $MH(\theta, d\theta^*)$ (MH représentant le ratio Metropolis entre $p(D \mid \theta^*, M)p(\theta^*)$ et $p(D \mid \theta, M)p(\theta)$ corrigé par le ratio Hasting des probabilités de proposition de mouvements entre $p(\theta^* \rightarrow \theta)$ et $p(\theta \rightarrow \theta^*)$). Combiné avec la méthode MH, le *Gibbs sampling* permet d'accélerer la convergence des chaînes MCMC étant donné que chaque nouvelle valeur de

paramètre proposée est tirée directement de sa distribution *a posteriori* conditionnelle à la valeur des autres paramètres. Mais, puisque ces probabilités postérieures conditionnelles sont habituellement complexes à calculer pour les paramètres continus, la performance de cet algorithme est préférentiellement mise à contribution pour l'échantillonnage de valeurs discrètes.

La qualité d'une chaîne MCMC dépend directement de la configuration du noyau stochastique autour duquel s'articulent les mouvements proposés. Les premiers cycles du MCMC traversent initialement une zone dite de *burn-in* avant de converger vers celles de la distribution postérieure ciblée. Après plusieurs cycles dans la zone de convergence, si elle devient ergodique et apériodique (Norris 1998), elle visitera les différents états avec une fréquence proportionnelle à leur probabilité postérieure respective. Un échantillon représentatif de la distribution postérieure pourra être extrait de sa trace si son temps de décorrélation (Norris 1998) est suffisamment court.

1.3 Les Modèles d'Évolution Moléculaire

1.3.1 Les modèles nucléotidiques classiques

Un modèle d'évolution moléculaire est essentiellement un processus qui décrit comment les nucléotides ou les acides aminés sont substitués le long d'une phylogénie. Et il assume que cet enchaînement de substitutions obéit à un processus Markovien continu en temps charactérisé par une matrice Q de taux de transition instantanés. Un modèle d'évolution nucléotidique serait par exemple défini par la matrice

$$Q = \begin{pmatrix} - & Q_{AC} & Q_{AG} & Q_{AT} \\ Q_{CA} & - & Q_{CG} & Q_{CT} \\ Q_{GA} & Q_{GC} & - & Q_{GT} \\ Q_{TA} & Q_{TC} & Q_{TG} & - \end{pmatrix} \tag{3}$$

Les entrées de sa diagonale étant telles que la somme de chacune de ses lignes soit égale à 0. Notons que sous un tel processus, le taux de transition au temps t_x ne dépend que de l'état courant. Il est totalement indépendant de l'histoire substitutionnelle antérieure.

Suivant un processus de Poisson (Johnson & Kotz 1969 ; Haight 1967), il est possible à partir de Q de simuler l'histoire substitutionnelle d'un site partant de la racine d'un arbre. Sous un processus Markovien continu en temps, le temps d'attente (t) avant la prochaine substitution partant d'un état a est tiré d'une distribution exponentielle de taux $-Q_{aa}$;

$$t \sim Exp(-Q_{aa}). \tag{4}$$

Sur un interval de temps total l, une telle simulation, partant d'un état i, répétée tant que la somme des temps t n'atteint pas au moins l, se termine dans l'état j avec une probabilité donnée par la matrice P de probabilité de transition tel que

$$P_{ij}(l) = [e^{lQ}]_{ij}. \tag{5}$$

Les processus de substitution Markoviens continus et réversibles en temps respectent l'équation du bilan détaillé suivant

$$\pi_i Q_{ij} = \pi_j Q_{ji}, \tag{6}$$

où π_i et π_j correspondent respectivement à la fréquence d'équilibre de l'état i et à celle de l'état j. Il stipule que la proportion de transitions allant de i vers j est la même que celle allant de j vers i. Cette hypothèse de réversibilité est effectivement pratique sur le plan technique mais n'est en aucun cas supportée par les données biologiques.

Les premiers modèles classiques réversibles en temps appliqués à l'évolution des séquences nucléotidiques étaient tous homogènes en sites et en temps. À l'époque, leur mode de paramétrisation s'est progressivement adapté aux nouvelles observations extraites des données biologiques empiriques. Par exemple, le modèle le plus simpliste proposé, JC69 (Jukes & Cantor 1969), assumait une uniformité des taux de substitution entre nucléotides. Hors, il a clairement été démontré qu'il exite un biais mutationnel favorisant les transitions (particulièrement C vers T) par rapport aux tranversions. Dans le cas des séquences codantes, les transitions sont également plus fréquentes puisqu'elles correspondent le plus souvent à des substitutions d'acides aminés qui sont conservatrices. Les transversions impliquent quant à elles davantage des substitutions entre des acides aminés biochimiquement différents et donc moins bien tolérées. Le modèle K80 (Kimura 1980) a alors été proposé pour accommoder cette réalité évolutive. Celui-ci contient un paramètre supplémentaire, κ, qui désigne le taux relatif de transition par rapport au taux de transversion. Il est en pratique de l'ordre de 2 sur des données nucléaires mais peut facilement atteindre 5 ou plus sur du mitochondrial métazoaire.

Les modèles de JC69 et K80 assument tous deux que les 4 nucléotides ont la même fréquence d'occurrence dans le génome. Les génomes mitochondriaux des métazoaires, riches en AT, est un bel exemple d'incompatibilité avec ce type d'uniformité. Ce qui explique le besoin de concevoir un modèle tel celui de Felsenstein (1981) connu sous le nom de F81. Il suggère que durant le processus substitutionnel Markovien, les 4 nucléotides ont des fréquences d'équilibre respectives (π_i) qui ne sont pas nécessairement les mêmes et que ce sont uniquement elles qui définissent les taux de transition dans la matrice Q ($Q_{ij} = \pi_j \quad \forall i$). Le modèle HKY85 (Hasegawa et al. 1985) propose quant à lui d'accommoder simultanément la non-uniformité des fréquences d'équilibre et le taux relatif de transition par rapport au taux de transversion en combinant les modèles F81 et K80.

Finalement, le modèle d'évolution nucléotidique le plus général, homogène et réversible en temps (GTR), a été conçu par Tavaré (1986). Il remplace le paramètre κ du modèle HKY85 par 6 taux relatifs d'échange symétriques (ρ_{ij}) entre les 4 nucléotides. Il est ainsi possible par ML ou par inférence Bayésienne d'estimer librement des ρ_{ij} et des π_i directement à partir du jeu de données empriques. La propriété $\rho_{ij} = \rho_{ji}$ du modèle GTR est ce qui lui permet de préserver la réversibilité en temps du processus substitutionnel selon le bilan détaillé suivant ;

$$\pi_i \rho_{ij} \pi_j = \pi_j \rho_{ji} \pi_i. \tag{7}$$

1.3.2 Les modèles d'évolution pour les protéines

Le processus Markovien continu et réversible en temps que nous venons de décrire pour les acides nucléiques s'applique également aux protéines. Mais puisque qu'elles sont constituées de 20 acides aminés, il est souvent plus pratique de modéliser leur processus de substitution GTR avec des matrices empiriques plutôt que de l'inférer directement à partir du jeu de données étudié.

Margaret Dayhoff et ses collaborateurs (Dayhoff et al. 1972 ; Dayhoff et al. 1978) furent les premiers à proposer une matrice 20×20 de taux substitution entre acides aminés connu sous le nom de modèle de Dayhoff. Cette matrice résulte de l'observation de 71 jeux de données constitués de protéines peu divergentes ($> 85\%$ en similarité). Chacune de ses entrées correspond à une probabilité de transition " PAM1 " (*Point Accepted Mutation*) calculée à partir de 1 572 substitutions inférées parcimonieusement. Il est possible de convertir une matrice PAM1 en un modèle de probabilité de transition réflétant des temps de divergence plus longs. Par exemple, la matrice PAM250 résulte du calcul de PAM1 élevée à la puissance 250.

À l'époque, relativement peu de données protéiques étaient disponibles pour concevoir le modèle de Dayhoff. Mais plus tard, avec l'accumulation progressive des données et grâce aux travaux exécutés par Jones et al. (1992), un modèle équivalent, construit à partir de jeux de données beaucoup plus larges, à vu le jour ; les entrées de la matrice PAM1 JTT réflètent 59 190 substitutions inférées parcimonieusement à partir de jeux de données nucléaires totalisant 16 130 protéines. À la base, ce modèle n'est pas compatible avec l'hypothèse du processus de substitution Markovien continu et réversible en temps. Pour l'utiliser en inférence probabiliste, il est nécessaire de convertir, par estimation mathématique (Nilesen 2005), les matrices Dayhoff et JTT sous forme de matrices $\mathbf{R}$ de taux relatifs symétriques. Les matrices $\mathbf{R}$ sont communément connues sous le nom de modèle REV.

L'approche utilisée pour concevoir les modèles empiriques Dayhoff et JTT est relativement rapide mais se limite à l'information extraite de séquences protéiques peu divergentes. Les approches probabilistes sont cependant mieux adaptées pour tenir compte de la probabilité élevée de substitutions multiples dans le cas des jeux de données plus divergents. C'est pourquoi Adachi & Hasegawa (1996) ont utilisé la méthode ML pour développer le modèle mtREV à partir de génomes mitochondriaux complets provenant de 20 espèces de vertébrés. Le modèle résultant a ainsi été estimé avec un degré de liberté de 208 ; 190 paramètres Q_{ij} et 19 fréquences d'équilibre associés aux 20 acides aminés. Les protéines codées par les génomes mitochondriaux animaux étant majoritairement transmenbranaires hydrophobes, elles évoluent sous différentes contraintes sélectives. La matrice mtREV est par conséquent non recommandée pour tenter d'expliquer l'histoire évolutive de jeux de données constitués de protéines nucléaires (essentiellement hydrophiles). L'équivalente ML nucléaire est la matrice WAG (Whelan & Goldman 2001) élaborées à partir de 3 905 protéines divisées en 182 familles. Des tests de ratio de vraisemblance (LRT) (Neyman &

Pearson 1933) ont révélé que les modèles issus de l'estimation ML sont plus performants que les modèles Dayhoff et JTT.

Il est bien connu, que les sites à l'intérieur d'une même protéine évoluent sous différentes contraintes sélectives. Ceux soumis à une forte pression sélective sont dits lents ou invariants contrairement à ceux dits rapides qui sont soumis à peu ou pas de pression sélective. La matrice LG (Le & Gascuel 2008) est une extension de la matrice WAG permettant d'accomomder cette réalité évolutive. Élaboré à partir de jeux de données plus larges et plus diversifiés, le modèle LG semble mieux ajusté à plusieurs jeux de données comparativement à WAG. De plus, LG s'est révélé plus juste quant à la reconstruction de certaines topologies.

Nous aborderons plus en détail l'hétérogénéité quantitative (taux de substitution à travers les sites) et qualitative (mode de substitution à travers les sites) des processus de substitution. Mais pour l'instant mentionnons qu'il existe également des matrices empiriques estimées à partir de différentes catégories de sites partionnés dépendamment de leur structure secondaire et de leur accessibilité à l'environnement acqueux (Koshi & Goldstein 1995, Thorne et al. 1996a, Goldman et al. 1998). Plus généralement, des modèles de mélange empiriques combinant simultanément un ensemble de matrices de transition élaborées dans un contexte de ML ont été proposés (Quang et al. 2008b). Les mélanges de matrices sont estimés à partir de jeux de données extraits de la banque de données HSSP (Schneider et al. 1997) avec une approche supervisée (EX2, EX3, EHO) ou non-supervisée (UL2 et UL3). Et chacune des matrices d'un mélange spécifique est compatible avec un ensemble de sites plus ou moins exposés à l'environnement acqueux ou structurellement différents. Globalement, les modèles de mélange empiriques sont mieux ajustés à divers jeux de données nucléaires comparativement aux modèles constitués d'une seule matrice de transition (JTT, WAG et LG).

Des modèles de mélange de matrices empiriques profil-spécifiques (Quang et al. 2008a, Wang et al. 2008) ont également été estimés. Par exemple, dans le cas de Quang et al. (2008a), les constituants de ces mélanges ont été estimés à partir de jeux de données extraits de la banque de données HSSP avec un algorithme d'espérance-maximisation (EM) (Dempster et al. 1977). Les mélanges sont constitués de 10 à 60 matrices (C10, C20, C30, C40, C50 et C60) différentes entre elles uniquement de par leur vecteur de fréquence d'équilibre en acides aminés (profil). Chacun tente de modéliser les propensions différentes qu'ont les sites pour les 20 acides aminés dépendamment de leur contexte environnemental respectif. Selon les résultats de tests de performance obtenus, le modèle C20 serait suffisant pour permettre un meilleur ajustement aux données comparativement au modèle WAG. De tels mélanges de 3 et 6 matrices empiriques (ECG3 et ECG6) ont également été estimé avec l'algorithme EM par Groussin et Lartillot (en préparation) mais à partir de jeux de données nucléaires différents. Les taux relatifs d'échange entre acides aminés et les profils du mélange ECG6 sont d'ailleurs ceux avec lesquels nous avons travaillé tout au long de ce projet.

1.3.3 Les modèles non-homogènes entre sites

Nous avons brièvement introduit le phénomène de variation des taux d'évolution à travers les sites avec le modèle LG. Rappelons qu'il est dû au fait que la contrainte sélective agissant n'est pas homogène le long des séquences nucléotidiques et protéiques. Par exemples, aux niveaux des régions codantes, les codons tolèrent préférentiellement des substitutions en troisième position étant donné qu'elles sont le plus souvent synonymes. Fitch & Margoliash (1967) furent les premiers à observer une telle hétérogénéité évolutive sur le cytochrome c. Les ARN fonctionnels et de structure ainsi que les protéines comportent certains sites localisés dans des régions stratégiques. Ceux-ci évoluent forcément moins rapidement que les sites situés à l'extérieur de telles régions.

Il est bien connu que négliger de tenir compte de ce type d'hétérogénéité substitutionnelle peut avoir des conséquences sérieuses sur les topologies et les temps de divergences estimés. Il est alors possible d'atténuer de tels effets en supprimant les sites rapides (Brinkmann & Philippe 1999). Cependant, cette approche peut entraîner la suppression de sites porteurs de signaux phylogénétiques importants. Yang (1993) introduit une méthode beaucoup plus robuste consistant à assumer que la vitesse d'évolution à chacun des sites est une variable aléatoire tirée d'une distribution gamma (Γ). Ce modèle connu sous le nom de RAS (*Rate Across Sites*) permet d'augmenter considérablement la vraisemblance des phylogénies inférées. Toutefois, sous ce modèle il est nécessaire d'intégrer la vraisemblance conditionnelle sur l'ensemble des taux à chacun des sites. C'est la raison pour laquelle il est beaucoup plus pratique d'approximer cette distribution en la discrétisant (D-Gam) en m catégories (Yang 1994). Quatre catégories discrétisées est souvent suffisant pour un ajustement optimal aux données. Le calcul de la vraisemblance conditionnelle est ainsi largement simplifié en sommant sur les quatres taux correspondant respectivement à la médiane de chacune des catégories. Le paramètre de forme (α) de la D-Gam est un paramètre libre du modèle qui peut être estimé par ML ou par inférence Bayésienne directement à partir des données. Celui-ci est fréquemment combiné avec un second paramètre permettant d'estimer la proportion de sites invariant (I) pour lesquels $r = 0$ (Gu et al. 1995).

Accommoder dans les modèles phylogénétiques le fait que les sites évoluent plus ou moins rapidement est effectivement essentiel. Mais si l'on y réfléchit bien, la dimension qualitative des processus de substitution devrait tout autant être hétérogène entre les sites. Dépendamment de leur emplacement ponctuel dans la protéine et des propriétés physico-chimiques de leur environnement immédiat, certains sites devraient tolérer plus ou moins bien la présence de certains acides aminés.

Thorne et al. (1996b) ont introduit un modèle empirique probabiliste qui relie la structure secondaire (hélice-α, feuillet-β, boucle) des protéines à des processus évolutifs distincts. Avec une approche similaire à celle adoptée Dayhoff et al. (1972), ils ont effectivement pu observer que les taux relatifs d'échange entre acides aminés ainsi que les fréquences d'équilibre de chacun sont différents selon l'alignement de structures secondaires connues

utilisé pour les estimer. Koshi & Goldstein (1995) ont déjà également tenté une telle approche mais ont reconnu le manque d'hétérogénéité d'un tel modèle. C'est que celui-ci assume toujours que les sites, qu'ils soient par exemple les constituants d'une hélice-α dans un environnement hydrophobe ou hydrophile, continuent d'évoluer sous le même régime d'évolution.

Un modèle de mélange assume l'existence d'une nombre fini ou infini de composantes (catégories ou classes) définissant un processus évolutif spécifique. Dans ce que nous avons vu jusqu'à maintenant, les taux de transition caractérisant une matrice sont fonction de l'identité des acides aminés. Koshi et Goldstein (1998, 2001) adoptent quant à eux une approche basée sur les propriétés physico-chimiques (par exemple, l'hydrophobicité et la longueur de la chaîne latérale) des acides aminés au lieu de leur identité. Les classes S_k constituant les mélanges de leur modèle sont ici en fait des fonctions linéaires ou quadratiques évaluant respectivement la valeur de compatibilité d'un acide aminé à un site donné dépendamment de ses propriétés physico-chimiques. De plus, contrairement au modèle RAS, les vitesses d'évolution sont classe-dépendantes. Les taux de transition définis dans les différentes matrices Q^k sont soumis à une cinétique de Métropolis ; c'est-à-dire que le taux Q^k_{ij} dépend de s'il y a un gain ou une perte de valeur de compatibilité dans cette classe S_k selon les propriétés physico-chimiques des acides aminés i et j. Cette méthode permet entre autres de réduire considérablement le nombre de paramètres libres du modèle. Une série de mélanges de classes (dont le nombre varie de 2 à 11) ont été estimés par EM avec cette approche à partir d'un jeu de données constitué de protéines d'enveloppe du virus HIV. Des tests de performance ont démontré que ces modèles de mélanges sont mieux ajustés aux données que les modèles avec une seule matrice de substitution. Ce qui encore ici met en évidence l'importance de prendre en compte l'hétérogénéité qualitative site-spécifique des processus de substitution.

Les différents modèles de mélange empiriques que nous avons décrits présupposent un degré d'hétérogénéité fixé *a priori* et peu élevé. Bruno (1996) et Halpern & Bruno (1998) poussent à l'extrême ce degré d'hétérogénéité ; chacun des sites évoluant sous processus de substitution profil-spécifique différent. Le désavantage d'un tel modèle est qu'il requiert un nombre important de taxons pour que les fréquences d'équilibre estimées individuellement à chacun des sites soient statistiquement significatives. Aucun des modèles n'avaient donc tenter d'extraire à même les données observées l'étendue réel d'un tel phénomène.

Comme nous l'avons déjà mentionné, l'inférence Bayésienne est beaucoup plus flexible par rapport aux méthodes ML quant à son degré de paramétrisation potentiel. Lartillot & Philippe (2004), avec la méthode MCMC, ont utilisé une approche non-paramétrique pour tenter d'évaluer *a posteriori* le degré d'hétérogénéité spatiale qualitative des processus de substitution à partir de différents jeux de données. Ce modèle connu sous le nom de CAT (pour **cat**égories de profil), selon qu'il utilise un mélange de profils empiriques ou non, peut être paramétrique ou non-paramétrique. Contrairement aux modèles non-paramétriques, les modèles paramétriques assument que la loi définissant la distribution

d'une caractéristique site-spécifique est connue *a priori*. Le modèle RAS est un exemple de modèle qui assume que les vitesses de substitution à travers les sites obéissent à une distribution gamma.

Sous la configuration non-paramétrique du modèle CAT, le nombre de profils site-spécifiques est non fixé *a priori*. On parle alors d'un mélange infini de profils tirés de manière indépendante et identiquement distribués d'une loi dont la forme est inconnue. Celle-ci est en fait inférée directement à partir des données avec un processus de Dirichlet (Ferguson 1973). Globalement, les résultats obtenus *a posteriori* supportent un degré d'hétérogénéité supérieur à ceux des modèles de mélange proposés par Koshi & Goldstein (1998) et Thorne et al. (1996b). De plus, des tests de performance ont révélé que dans tous les cas CAT semble mieux ajusté aux données observées comparativement aux modèles homogènes et au modèle de Bruno (1996) introduit précédemment.

Un cas intéressant que nous aimerions présenter ici, mettant en évidence la performance du modèle CAT, implique deux groupes de métazoaires reconnus pour évoluer rapidement ; les nématodes et les platyhelminthes. Lartillot et al. (2007) ont démontré que leur modèle a effectivement la capacité de résister à l'artéfact LBA en positionnant correctement les nématodes et les platyhelminthes dans un arbre incluant des deutérostomes et des arthropodes, enraciné avec un groupe de champignons. Sous un modèle homogène entre sites tel WAG, l'attraction biaisée des nématodes et des platyhelminthes vers celui du outgroup doit être rompue par l'ajout de taxons supplémentaires (Choanoflagellés et Cnidaires) à la base de ce outgroup (figure 2). Ce qui n'est pas nécessaire sous le modèle CAT, qui lui positionne les nématodes et les platyhelminthes à l'intérieur des protostomes (i.e. avec les arthropodes) indépendamment de l'enrichissement ou non du outgroup (figure 3).

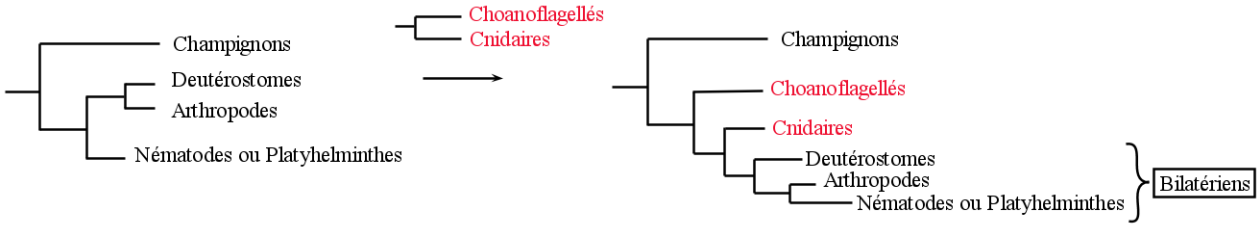


Figure 2 – Sensibilité du modèle WAG à l'artéfact LBA. Un meilleur échantillonnage de taxons est nécessaire pour récupérer la monophylie des protostomes (nématodes + arthropodes). Figure modifiée, d'après Lartillot et al. (2007).

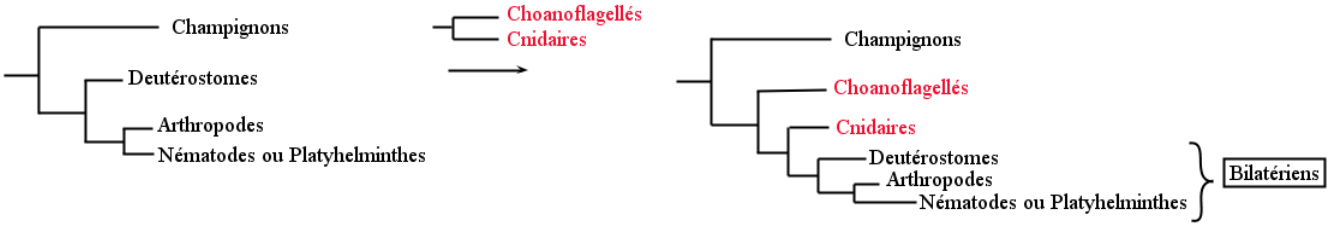


Figure 3 – Résistance du modèle CAT à l'artéfact LBA. La monophylie des protostomes est détectée avant même la rupture de la longue branche séparant les nématodes des champignons. Figure modifiée, d'après Lartillot et al. (2007).

1.3.4 Les modèles non-homogènes en temps

Nous venons d'aborder la problématique phylogénétique liée au fait que les sites n'évoluent pas nécessairement tous à la même vitesse et peuvent avoir des propensions différentes pour les 20 acides aminés. Ce sont deux caractéristiques évolutives qui dépendent à la fois de la force et de la nature des contraintes sélectives qui agissent sur eux. Mais s'ajoute à cette hétérogénéité entre sites, une hétérogénéité en temps. C'est à dire une dérive compositionnelle qui modifie le contenu en guanine + cytosine (G+C) des génomes à travers le temps. Cette dérive peut se répercuter sur le contenu en acides aminés dépendamment si les positions non-synonymes des codons sont ciblées par un tel biais.

Le contenu G+C peut varier considérablement d'une espèce à l'autre. Chez les bactéries, il peut varier entre 25% et 75% (Muto & Osawa 1987). Cette variation est également observable au niveau génique. C'est le cas par exemple du cytochrome b mitochondrial (Jermiin et al. 1994). L'impact qu'a cette réalité évolutive sur les reconstructions phylogénétiques est non négligeable. Des espèces éloignées aux contenus G+C similaires peuvent être incorrectement regroupées ensemble. Une alternative souvent adoptée est alors d'atténuer les effets du biais compositionnel génomique en utilisant des jeux de données protéiques. Mais Foster & Hickey (1998) ont démontré que cette stratégie n'est pas nécessairement toujours à l'abris de ce phénomène. Par exemple, avec des jeux de données mitochondriaux géniques et protéiques animal, différentes approches basées sur des modèles homogènes en temps regroupent systématiquement les abeilles et les nématodes. Cet artéfact est effectivement causé par leurs contenus A+T similaires (riches) ciblant les positions non-synonymes des codons et qui se répercutent sur leurs contenus en acides aminés (riches en FYMINK). Soulignons cependant que les biais compositionnels observés aux niveaux des protéines ne sont pas uniquement une conséquence de ceux observés au niveau nucléotidique. Ils peuvent également réfléter une adaptation spécifique de l'organisme à l'environnement dans lequel il évolue (Bogatyreva et al. 2006, Das et al. 2006).

Si l'on observe aujourd'hui de telles différences dans les proportions G+C versus A+T entre différents génomes, c'est que nécessairement ces biais de composition se sont ma-

nifestés ancestralement à travers l'évolution des espèces. Donc, la modélisation d'un processus de substitution homogène à travers le temps n'est plus valide. Différentes approches temps-hétérogènes ont été tentées pour simuler la dérive compositionnelle et atténuer son impact sur les reconstructions phylogénétiques (Foster 2004, Galtier & Gouy 1998, Blanquart & Lartillot 2006, Blanquart & Lartillot 2008).

Galtier & Gouy (1998) ont contribué à mettre en évidence la dérive compositionnelle du contenu G+C à travers le temps en introduisant dans leur modèle un paramètre (π_{GC}) effet par branche. Chaque branche j a la possibilité d'ajuster son contenu G+C en maximisant la vraisemblance de son π^j_{GC} sachant les données observées et le modèle homogène de Tamuara (1992) suivant ;

$$Q = \begin{pmatrix} - & \pi^j_{GC}/2 & \kappa\pi^j_{GC}/2 & (1-\pi^j_{GC})/2 \\ (1-\pi^j_{GC})/2 & - & \pi^j_{GC}/2 & \kappa(1-\pi^j_{GC})/2 \\ \kappa(1-\pi^j_{GC})/2 & \pi^j_{GC}/2 & - & (1-\pi^j_{GC})/2 \\ (1-\pi^j_{GC})/2 & \kappa\pi^j_{GC}/2 & \pi^j_{GC}/2 & - \end{pmatrix} \tag{8}$$

où κ, le taux de transition versus taux de transversion, est uniforme à travers toutes les branches. Leur modèle évalué avec divers tests de simulation inférentielle réussit à estimer correctement les différentes valeurs de π^j_{GC}. Et ces π^j_{GC} semblent effectivement hétérogènes à travers le temps.

Le nombre élevé de paramètres π^j_{GC} peut considérablement réduire la performance du modèle de Galtier & Gouy (1998). Foster (2004) proposa donc une approche Bayésienne pour réduire ce risque de surparamétrisation. Au lieu d'estimer la valeur d'un paramètre π_{GC} pour chacune des branches, son modèle combiné à la méthode MCMC permet de s'ajuster aux données en ajoutant ou en supprimant des vecteurs profil-spéciques sur différents noeuds à travers l'arbre. Le nombre de vecteurs, leur profil respectif ainsi que leur emplacement sont estimés *a posteriori*. Ce modèle a démontré qu'il est mieux ajusté aux données observées que les modèles homogènes en temps et qu'il peut résister aux artéfacts de reconstruction causés par les biais compositionnels.

À bien y réfléchir, comment les changements dans les contenus G+C peuvent-ils être dépendants des événements de spéciation. C'est pourtant ce que suggère les modèles de Galtier & Gouy (1998) et de Foster (2004) en réévaluant le contenu G+C aux noeuds de l'arbre. Blanquart & Lartillot (2006) introduisirent un modèle *Break Point* (BP) autorisant le découplage entre ces deux événements. Avec une approche Bayésienne similaire à celle de Foster (2004), leur modèle permet d'échantillonner de la distribution postérieure conditionnelle aux données observées des BP correspondant à des changements de profils le long de l'arbre. La principale différence avec le modèle de Foster (2004) est que ces BP sont des événements ponctuels apparaissant le long des branches (figure 4) et non pas aux niveaux des noeuds. Une branche peut donc être ciblée par 0 ou plusieurs BP. Le modèle permet aussi de résoudre des phylogénies incorrectement inférées sous des modèles homogènes en temps dû au biais de composition G+C. De plus, contrairement au modèle

de Galtier & Gouy (1998), il est en accord avec la réalité biologique voulant qu'un même profil puisse être conservé sur plusieurs branches consécutives. Appliqué sur différents jeux de données d'ARNr 16S bactériens, il infère *a posteriori* beaucoup moins de BP que le modèle de Galtier & Gouy (1998). De par sa plus grande flexibilité, il est également mieux ajusté aux données dont le contenu G+C est peu hétérogène ; le nombre de BP estimé *a posteriori* sous de tel jeux est près de 0.

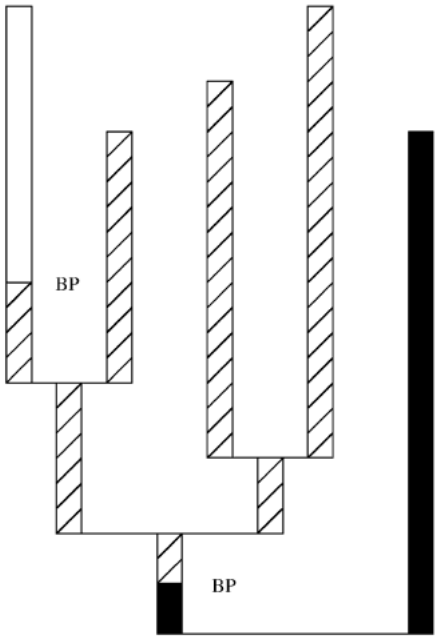


Figure 4 – Exemple d'une réalisation sous le modèle BP de Blanquart & Lartillot (2006). Deux BP sont placés sur deux branches différentes ; l'un définissant la région hachurée et l'autre la région blanche. Figure tirée de Blanquart & Lartillot (2006).

1.3.5 Les modèles Markov-modulés pour les taux d'évolutions

Le modèle RAS assume que le taux d'évolution à un site donné tiré de la D-Gam est homogène à travers l'ensemble des branches d'un arbre. C'est l'hypothèse de l'homotachie (terme Grecque signifiant " même vitesse "). D'autre part, les travaux de Fitch (1971a) sur le cytochrome c ont démontré par analyse phylogénétique que l'ensemble des codons invariables chez les champignons n'est pas le même que celui chez les métazoaires. Ce qui serait plutôt en accord avec l'hypothèse inverse qu'est l'hétérotachie. C'est le terme général employé en évolution moléculaire pour désigner les variations de taux d'évolution site-spécifiques au cours du temps. Les observations de Fitch (1971a) sur l'évolution hétérogène entre taxons du cytochrome c font cependant plus précisément référence au phénomène covarion (Fitch & Markowitz 1970). Cette hypothèse hétérotache stipule qu'une proportion spécifique de codons varient de façon concomitante alors que l'autre proportion est invariable. Les codons invariables seraient en fait ceux qui sont responsables de coder pour des acides aminés essentiels au maintien de la fonction de la protéines. Et suite à un

déplacement spatial de l'ensemble des contraintes sélectives agissant sur la protéine, les codons variables et invariables ne seraient plus nécessairement les mêmes.

Il a été démontré que le phénomène d'hétérotachie est très répandu dans les jeux données utilisés pour les reconstructions phylogénétiques. Par exemple Lopez et al. (2002) ont observé que 95% des sites variables sont hétérotaches sur un jeu de données protéique de cytochrome *b* vertébré. Omettre d'accommoder l'hétérotachie dans les modèles d'évolution moléculaire peut potentiellement mener à des reconstructions fausses (Philippe et al. 2005, Lockhart et al. 1996, Inagaki et al. 2004, Kolaczkowski & Thornton 2004).

Tuffley & Steel (1998) furent les premiers a tenter d'accommoder le processus covarion avec un modèle Markov-modulé. Sous ce modèle, les sites peuvent moduler entre deux états cachés le long des branches. L'un nommé " OFF " (invariable) et l'autre " ON " (variable). Le processus de substitution de l'état ON est conforme au processus Markovien continu et réversible en temps et ce sous un seul et même taux d'évolution à travers les sites et le temps. Le taux de transition de l'état OFF vers l'état ON et de l'état ON vers l'état OFF sont respectivement S_{01} et S_{10}. La matrice Q Markov-modulée résultante combinant les deux processus (transitions entre états cachés et transitions entre états observés) est d'ordre 4×2 dans un cas nucléotidique et d'ordre 20×2 dans un cas protéique.

Afin d'accommoder simultanément l'hétérogénéité des taux d'évolution à travers les sites, Huelsenbeck (2002) adopte une stratégie légèrement différente de celle proposée par Tuffley & Steel (1998). Au lieu d'un modèle avec une seule matrice Q Markov-modulée, il opte pour un modèle comportant autant de matrices Q que le nombre (m) de catégories de la D-Gam. Chacun des sites module indépendamment entre un état ON, dont le taux est tiré de la D-Gam, et un état OFF.

Le modèle de Galtier (2001) revient à un seul processus Markov-modulé à travers les sites modulants mais comporte g états cachés égal au nombre de catégories discrètes de la D-Gam. Chacun des m taux médians de la D-Gam est attribué à l'un des états cachés. Donc, l'état OFF n'est pas intégrer à ce modèle. Et contrairement aux modèles ON/OFF, les taux de transition entre états cachés sont uniformes ; un seul taux égal à S_{11}/g. De plus, une proportion de sites sont exclus du processus modulant ; ceux-ci évoluent sous un régime homotache avec chacun leur propre taux d'évolution tiré de la D-Gam.

Nous avons pu noter que les modèles de Huelsenbeck (2002) et de Galtier (2001) combinent hétérotachie et RAS de façon complètement différente. Wang et al. (2007) proposent de combiner les deux dans un modèle covarion généralisé (figure 5). Celui-ci autorise aussi une certaine proportion de sites non-modulants à la manière de Galtier (2001). Tous les autres sites évoluent sous le même processus Markov-modulé avec g états cachés ON et un état caché OFF. Les taux des états ON correspondent aux taux de la D-Gam tel que proposé par Galtier (2001). Ceux entre les états ON et l'état OFF (S_{01}, S_{10}) et ceux entre les états ON (S_{11}/g) sont modélisés par la matrice G présentée sur la figure 6.

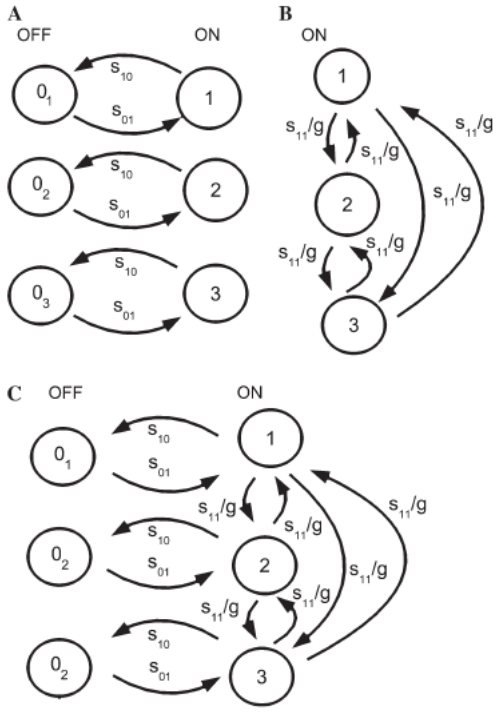


Figure 5 – (A) Modèle covarion de Huelsenbeck (2002). (B) Modèle covarion de Galtier (2001). (C) Modèle covarion généralisé de Wang et al. (2007). Figure tirée de Wang et al. (2007).

$$\mathbf{G} = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ g \\ 0_1 \\ 0_2 \\ \vdots \\ 0_g \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & \cdots & g & 0_1 & 0_2 & \cdots & 0_g \end{array} \\ \begin{pmatrix} * & s_{11}/g & \cdots & s_{11}/g & s_{10} & 0 & \cdots & 0 \\ s_{11}/g & * & \cdots & s_{11}/g & 0 & s_{10} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{11}/g & s_{11}/g & \cdots & * & 0 & 0 & \cdots & s_{10} \\ s_{01} & 0 & \cdots & 0 & * & 0 & \cdots & 0 \\ 0 & s_{01} & \cdots & 0 & 0 & * & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{01} & 0 & 0 & \cdots & * \end{pmatrix} \end{array}$$

Figure 6 – Modèle covarion généralisé de Wang et al. (2007) : matrice G de taux de transition entre l'état OFF et l'état ON (S_{01}, S_{10}) et entre les états ON (S_{11}/g). Figure tirée de Wang et al. (2007).

Wang et al. (2007) ont évalué la performance de leur modèle en comparant des valeurs d'ajustement obtenues avec 23 jeux de données différents par rapport à celles obtenues avec les modèles RAS, Huelsenbeck (2002) et Galtier (2001). Dans tous les cas, les modèles covarion performent mieux que RAS. Ce qui suggère effectivement le besoin d'accommoder le phénomène d'hétérotachie. Le ratio de meilleure performance entre les modèles de Huelsenbeck (2002) et de Galtier (2001) est de 16 :7. Ce partage de performance suggère quant

à lui le besoin de combiner les deux. Ce qui, en effet, est confirmé puisque le modèle de Wang et al. (2007) est significativement le mieux ajusté à 18 des 23 jeux de données.

1.3.6 Modéliser les changements temporels qualitatifs site-spécifiques dans les processus de substitution

Nous avons vu qu'il a clairement été démontré que les processus de substitution site-spécifiques sont quantitativement et qualitativement hétérogènes entre sites. Nous venons également de discuter à propos de l'importance d'accommoder les changements temporels dans les taux d'évolution site-spécifiques. Dans la suite, nous demeurerons dans cette dimension temporelle mais principalement dans une perspective qualitative. Nous introduirons en fait les changements de **mode** site-spécifiques dans les processus de substitution au cours du temps. Nous ferons référence au mode du processus de substitution comme étant la préférence marquée qu'a un site spécifique pour certains acides aminés ou nucléotides. Mais dans un autre contexte, ce mode pourrait également tout aussi bien faire référence par exemple à un taux de transition/transversion (κ) préférentiel.

Rappelons que la **force** d'une contrainte sélective agissant à un site donné est essentiellement ce qui détermine la vitesse d'évolution de ce site. Si par exemple un acide aminé particulier accomplit une fonction bien précise et fondamentale au sein d'une enzyme ou d'une protéine structurale, celui-ci aura une faible tendance à être substitué par un autre acide aminé. Dans le cas contraire, un site serait plus tolérant et pourrait plus aisément accommoder le passage d'un acide aminé vers un autre. La **nature** d'une contrainte fait quant à elle référence à la gamme d'acides aminés tolérés à un site donné. C'est à dire son profil. Celui-ci dépend du contexte environnemental dans lequel le site évolue.

Puisque l'intensité de la force de contrainte appliquée à un site donné peut changer au cours du temps, pourquoi la nature de cette contrainte devrait-elle demeurer fixe. Imaginons que des acides aminés voisins à un site dans l'espace tridimensionnelle aient été substitués par d'autres aux propriétés physico-chimiques différentes. Cela implique forcément une altération de contexte environnemental. Le processus de substitution ancestral précédant à ce site n'étant par exemple compatible qu'avec des acides aminés hydrophobes pourrait suite à cette altération environnementale ne tolérer qu'une gamme restreinte d'acides aminés hydrophiles.

De toute évidence, dire qu'un site particulier change de processus de substitution profil-spécifique au cours du temps revient à dire qu'il module vers un profil CAT différent. Les travaux de Roure & Philippe (2011) portant sur cette problématique ont d'ailleurs exploité l'approche CAT de Lartillot & Philippe (2004) pour mettre en évidence l'hétérogénéité temporelle qualitative des processus de substitution. Globalement, leur stratégie consistait à appliquer le modèle CAT sur un même jeu de données mais séparément pour deux groupes taxonomiques différents. Ensuite, pour chacun des sites, estimer leur probabilité d'être affiliés au même profil CAT dans les deux groupes avec score $\mathrm{PIP_n}$ (*Probability of*

Identical Profiles over n *clades*). Un PIP_n de 0 indiquant que le site est mieux décrit par un profil CAT spécifique dans le premier groupe et par un autre différent dans le second.

Pour désigner un changement de mode dans un processus de substitution site-spécifique au cours du temps, Roure & Philippe (2011) introduisirent le terme hétéropécillie. Avec leur méthode de score PIP, ils ont pu démontrer que ce phénomène est largement répandu dans les jeux de données empiriques mitochondriaux et nucléaires. De plus, ils ont également évalué que hétéropécillie et taux d'évolution sont fortement corrélés. Ce qui était en fait attendu puisqu'un site soumis à une forte pression sélective est contraint à réduire sont taux de substitution mais aussi sa diversité d'acides aminés tolérés et par conséquent la probabilité qu'il explore différents profils au cours du temps.

Roure & Philippe (2011) ont démontré qu'assumer qu'il n'existe pas de changement qualitatif temporel dans les processus de substitution peut potentiellement mener à des artéfacts de reconstruction phylogénétique. Le cas étudié concerne les liens évolutifs existants entre Porifera, Cnidaria et Bilateria (Deuterostomia+Protostomia). Les morphologistes ainsi que les jeux de données nucléaires supportent la monophylie des Eumetazoa (Cnidaria+Bilateria) mais celle-ci n'est pas récupérer avec les jeux de données mitochodriaux. Cet artefact de type LBA serait causé par une attraction du groupe Bilateria (évolution rapide) par le groupe distant (Choanoflagellata). La présence de sites hétéropécilles dans le jeu de données mitochodrial utilisé constituerait une violation au modèle CAT (celui-ci assume que chaque site conserve le même profil au cours du temps). Ce qui semble effectivement le cas selon la figure 7. Le retrait progressif des sites les plus hétéropécilles sous CAT+Γ_4 (i.e CAT combiné à RAS avec une D-Gam discrétisée en 4 catégories) permet de récupérer la monophylie du groupe Eumetazoa.

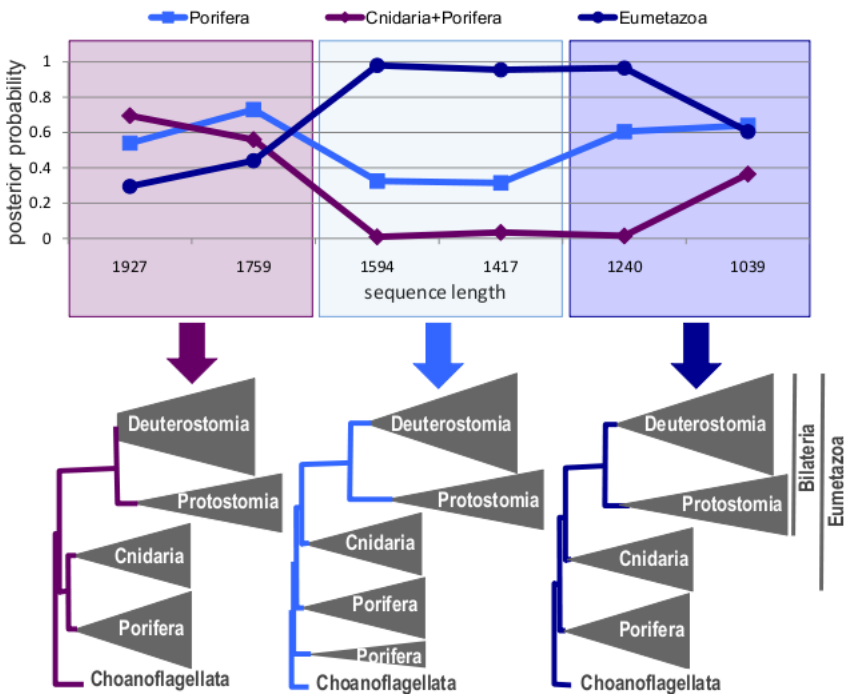


Figure 7 – Retrait progressif de sites hétéropécilles sous le modèle CAT+Γ_4 et récupération de la monophylie du groupe Eumetazoa. Figure tirée de Roure & Philippe (2011).

Des programmes informatiques de type classificateur existent et permettent de détecter des sites potentiellement hétéropécilles. L'idée de base de tels classificateurs est de classer les sites d'un alignement protéique empirique selon qu'ils aient été la cible de divergences fonctionnelles de type I (FDI) ou de type II (FDII) (Gu 1999). Ceux estimés comme ayant évolué sous la même contrainte sélective au cours du temps sont de type non-FD. Nous reviendrons plus loin sur le concept des divergences fonctionnelles qui sont causées soit par une duplication de gène, une spécialisation ou encore une perte de fonction protéique au cours de l'évolution (Henikoff & al. 1997, Li 1983). Mais ici, seulement pour préciser que les sites de type I sont essentiellement hétérotaches et ceux de type II sont considérés comme étant hétéropécilles.

Le programme FunDi (Gaston et al. 2011), selon des résultats de comparaisons fait partie des classificateurs les plus performants. Celui-ci utilise en premier lieu une approche similaire à celle de Roure & Philippe (2011) nécessitant de séparer horizontalement un jeu de données en deux sous groupes phylogénétiques (donc deux sous arbres). Il est un modèle de mélange à deux composantes ; la composante non-FD compatible avec les sites évoluant à travers l'arbre phylogénétique commun (les deux sous-arbres combinés) et la composante FD compatible avec ceux évoluant de manière indépendante à travers les deux sous-arbres. Avec une approche ML, les paramètres de taux d'évolution (lié au FDI) et de profil (lié au FDII) sont optimisés dans chacun des sous-arbres à tous les sites. Un score FD est attribué à chacun des sites dépendamment de la différence entre les valeurs de paramètres

estimées pour les deux sous-groupes.

La performance de différents classificateurs peut être mesurée à partir de jeux de données simulés incorporant des sites non-FD, FDI et FDII (Gaston et al. 2011). Il existe également des jeux de données empiriques de différents groupes de familles protéiques dont les sites ont été identifés expérimentalement comme étant de type non-FD, FDI ou FDII (Chakrabarti et al. 2007). Cependant, ces jeux de données ont l'inconvénient d'être reconnus comme étant " bruités" (présence potentielle de faux positifs et de faux négatifs). Gaston et al. 2011 ont utilisé les deux approches parallèlement à d'autres classificateurs et ont démontré que l'algorithme de FunDi semble le plus efficace.

1.3.7 Vers des modèles hétéropécilles Markov-modulés

Roure & Philippe (2011) et de Gaston et al. (2011) ont effectivement démontré que le phénomène d'hétéropécillie est bel et bien une réalité évolutive. Cependant, leurs approches ne contribuent pas concrètement à prendre en charge cette réalité en vue d'améliorer les reconstructions phylogénétiques. Holmes & Rubin (2002) proposent une approche prometteuse à cet égard basée sur la notion d'états cachés et les modèles de Markov-modulés. Les états cachés sous leur modèle sont caractérisés par des processus de substitution (matrices 20×20) différents de par leur profil de préférences en acides aminés entre lesquels les acides aminés sont autorisés à moduler au cours temps. Chacun des états observés au niveau de l'alignement peut ainsi potentiellement être dans un état caché au profil hydrophobe, polaire ou chargé.

Holmes & Rubin (2002) ont utilisé un algorithme EM pour estimer une série d'états cachés (1, 2, 3 et 4 états cachés) à partir de 200 alignements protéiques sélectionnés au hasard dans la base de données Pfam (Bateman et al. 2000). Un exemple avec deux états cachés est présenté à la figure 8. Notons qu'effectivement ces deux états cachés appris sont définis par deux processus de substitution opposés ; l'un (HS1) réflétant un processus de type hydrophobe et l'autre (HS2) de type hydrophile. L'application des quatres séries d'états cachés sur 200 autres alignements Pfam (jeux de données test) a permis de démontrer que l'augmentation du nombre d'états cachés augmente les valeurs de vraisemblance. Ce qui suggère qu'un nombre élevé d'états cachés est nécessaire pour modéliser adéquatement le processus évolutif des protéines au cours du temps. Cette augmentation du nombre d'états cachés semble d'ailleurs aussi contribuer à améliorer le niveau d'exactitude d'alignements protéiques multiples selon des tests de performance pratiqués sur des jeux de données BAliBASE (alignements de référence très précis obtenus à partir de structures crystallographiques)(Thompson et al. 1999).

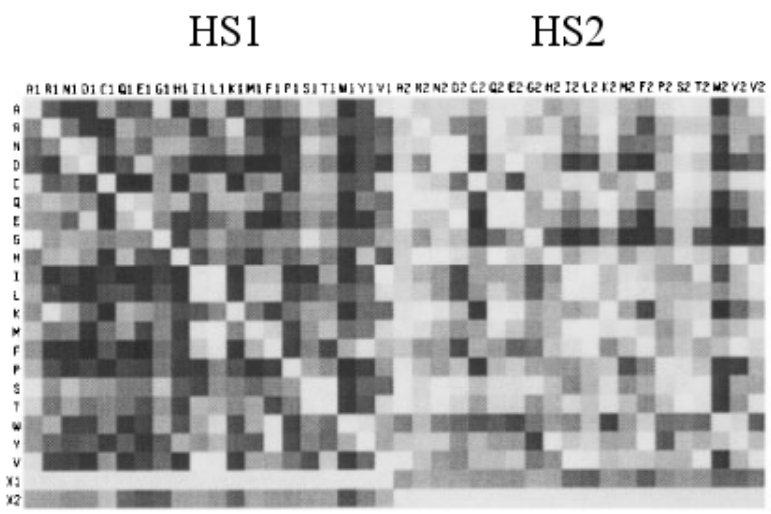


Figure 8 – Deux états cachés (HS1 et HS2) appris avec le modèle Markov-modulé de Holmes & Rubin (2002) sur des jeux de données sélectionnés de la base de donnés Pfam. Les carrés de couleur pâle indiquent un taux élevé de substitution. Les deux dernières lignes (X1 et X2) correspondent aux taux de transition entre les deux états cachés. Figure tirée de Holmes & Rubin (2002).

Whelan (2008) propose un modèle Markov-modulé pour les processus de substitution plus raffiné que celui de Holmes & Rubin (2002). Ce modèle semble optimisé pour les séquences nucléotidiques mais pourrait tout aussi bien être adapté aux séquences protéiques. Ce modèle est en fait une généralisation du modèle Markov-modulé covarion en autorisant tous les aspects du processus à varier au cours du temps. Il est constitué de G processus de substitution (états cachés) différents, chacun étant défini par son propre profil (π), taux de transition/transversion (κ) et taux d'évolution (μ). Chaque processus est donc un modèle HKY85 (Hasegawa et al. 1985) indépendant réversible en temps. La matrice de taux instantanés ($\mathbf{M}$) du kth processus est ainsi défini par

$$\mathbf{M}^k = \mu^k \begin{bmatrix} - & {\pi_C}^k & \kappa^k {\pi_G}^k & {\pi_T}^k \\ {\pi_A}^k & - & {\pi_G}^k & \kappa^k {\pi_T}^k \\ \kappa^k {\pi_A}^k & {\pi_C}^k & - & {\pi_T}^k \\ {\pi_A}^k & \kappa^k {\pi_C}^k & {\pi_G}^k & - \end{bmatrix}.$$

Contrairement aux modèles Markov-modulés covarion et au modèle de Holmes & Rubin (2002), celui de Whelan (2008) utilise une configuration GTR pour décrire le processus de transition entre les états cachés. Donc, les taux relatifs d'échange entre les états cachés ($\eth$) ainsi que les fréquences d'équilibre des états cachés (η) sont estimés à partir des données. L'ensemble des taux instantanés de transition entre les G états cachés est ainsi représenté

par la matrice GTR suivante ;

$$\mathbf{C} = \begin{bmatrix} - & \eth^{1,2}\eta^2 & \cdots & \rho^{1,G}\eta^G \\ \eth^{1,2}\eta^1 & - & & \eth^{2,G}\eta^G \\ \vdots & & \ddots & \\ \eth^{1,G}\eta^1 & \eth^{2,G}\eta^2 & & - \end{bmatrix}.$$

La combinaison de $\mathbf{M}$ et $\mathbf{C}$ produit le noyau stochastique Marko-modulé défini par

$$Q_{i,j}^{k,l} = \begin{cases} M_{i,j}^k & i \neq j, k = l \\ \pi_j^l C^{k,l} & i = j, k \neq l \\ 0 & i \neq j, k \neq l \end{cases}.$$

. Notons que le terme π_j^l signifie la fréquence d'équilibre de l'état observé j dans l'état caché l. Il est combiné avec $C^{k,l}$ pour assurer la réversibilité en temps du processus (i.e. satisfaire l'équation du bilan détaillé).

Une autre particularité du modèle de Whelan (2008) est que les taux d'évolution tirés de la D-Gam n'agissent pas seulement sur la vitesse globale des substitutions entre acides aminés mais également sur celle entre les états cachés. Donc, un site est rapide ou lent à la fois en terme de substitutions et en terme de modulations. Ce qui n'est pas le cas pour les modèles covarion décrits précédemment.

Lorsque les taux relatifs d'échange entre états cachés de la matrice $\mathbf{C}$ sont contraints à être tous égaux à 0 ($\eth^{k,l} = 0$), cela revient à un modèle de mélange homogène en temps. Et lorsque le nombre d'états cachés est de 1 ($G = 1$) et que RAS n'est pas incorporé, c'est un modèle homogène en temps et entre sites. Partant de cette idée, Whelan (2008) a pu évaluer la contribution des différents paramètres du modèle à l'hétérogénéité spatiale et temporelle des processus de substitution. Par exemple, en passant de 1 état caché vers un modèle de mélange à deux composantes différenciées seulement par π ($G = 2$, $\eth^{k,l} = 0$, $\pi^1 \neq \pi^2$), il est possible d'évaluer (par le gain en vraisemblance) la contribution de l'hétérogénéité des préférences en nucléotides à travers les sites. De même, en laissant libre les taux relatifs d'échange entre états cachés ($\eth^{GTR}$) et en augmentant le nombre d'états cachés, il est possible d'évaluer leur contribution à l'hétérogénéité temporelle par comparaison avec les modèles de mélange homologues (i.e vraisemblance sous $\eth^{k,l} = 0$ versus vraisemblance sous $\eth^{k,l} \neq 0$).

Cette approche choisie par Whelan (2008) a permis de mettre en évidence l'hétérogénéité temporelle et spatiale des processus de substitution à partir de différents jeux de données nucléotidiques. Globalement, augmenter du nombre d'états cachés et laisser libre les paramètres μ, π, κ et $\eth$ augmente toujours la vraisemblance. Mais, RAS semble être la forme

d'hétérogénéité la plus importante à tenir en compte. En fait, lorsque RAS n'est pas incorporé, les autres paramètres du modèle semblent tenter de décrire l'hétérogénéité des taux d'évolution à travers les sites par exemple avec 3 états cachés ayant des taux de transitions très différents (figure 9a). Notons également l'importance du phénomène d'hétérotachie à la figure 9a ; taux élevé de transition entre les états cachés 1 (évolution lente) et 3 (évolution rapide). En présence de RAS, notons à la figure 9b la redistribution de l'hétérogénéité temporelle et spatiale.

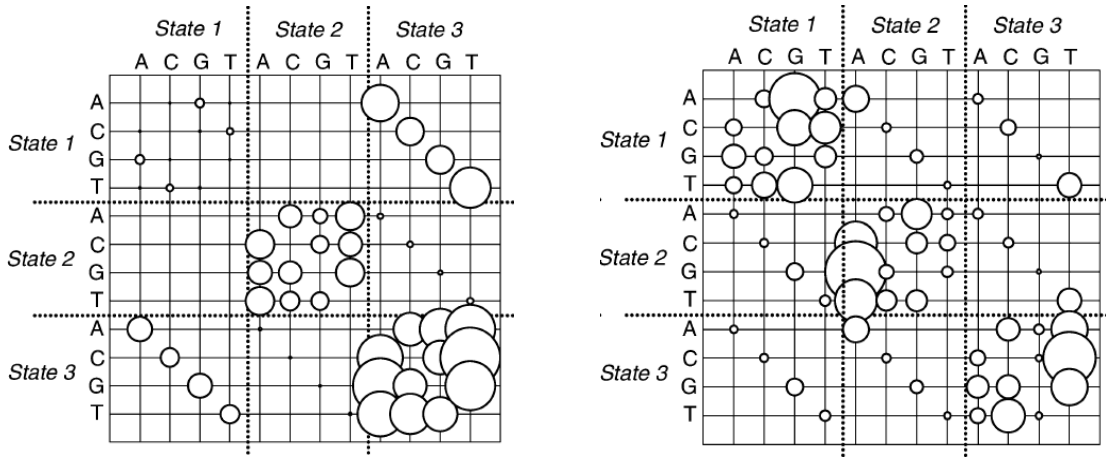


Figure 9 – *Bubble plot* illustrant les taux de transition de deux matrices Q du modèle Markov-modulé de Whelan (2008) estimées à partir d'un jeu de données constitué de 23 séquences nucléotidiques codantes d'entérobactéries. **a** Sans RAS. **b** Avec RAS (D-Gam avec 4 catégories). La taille des bulles est proportionnelle au taux de transition entre les deux états correspondants. Figures tirées de Whelan (2008).

Finalement, les résultats de Whelan (2008) avec seulement 3 états cachés, suggèrent fortement que des modèles d'évolution plus complexes sont nécessaires. La non-homogénéité qualitative et quantitative des processus de substitution, à travers les sites et à travers le temps, doit être accommodée pour reconstruire plus exactement des relations évolutives entre des organismes distants.

2 OBJECTIFS ET APPROCHES EXPÉRIMENTALES

L'objectif de ce projet de maîtrise consiste essentiellement à explorer plus en profondeur l'intérêt porté à l'égard des modèles Markov-modulés pour accommoder les changements de mode substitutionnel site-spécifiques au cours du temps. Notre travail partage ainsi des points communs avec la démarche introduite par Whelan (2008). Cependant, nous allons surtout nous focaliser sur la modélisation des processus substitutionnels appliqués aux protéines au lieu de nucléotidiques. Nous tenterons ainsi d'évaluer si un noyau stochastique Markov-modulé peut être approprié pour détecter dans des jeux de données protéiques des changements dans les préférences en acides aminés au cours du temps causés par exemple par des événements de divergences fonctionnelles (tel que thématisés par Gaston et al. (2011)) ou par des modifications de contextes physico-chimiques environnants.

Le protéome des microsporidies est reconnu pour contenir plusieurs sites ayant changé de régime substitutionnel au cours du temps comparativement aux autres champignons (Keeling & Fast 2002). C'est la raison pour laquelle nous avons choisi un jeu de données protéiques de microsporidies pour étudier notre modèle. Ces organismes unicellulaires parasites obligatoires ont suscité l'intérêt de la communauté scientifique pendant plus de 100 ans. Anciennement considérés comme les organismes eucaryotes les plus primitifs, diverses analyses de biologie moléculaire ont, au cours des dernières années, démontré qu'ils sont en fait des champignons (Keeling & Fast 2002). Tout indique que la lignée des microsporidies à évolué très rapidement et qu'elle a été la cible de sévères événements de miniaturisation sélective ayant touché leur structure cellulaire, leur métobolisme et leur matériel génétique.

De travailler avec des jeux de données constitués d'acides aminés (20 états observés) nous engage à faire face à de nouveaux défis. Il nous faut tout d'abord bien définir les profils (états cachés), c'est-à-dire les différents modes biochimiques entre lesquels vont moduler les sites au cours du temps. Pour ce faire, nous nous sommes appuyé sur des modèles de mélange empiriques préalablement estimés à partir de bases de données. D'autres part, l'augmentation considérable du nombre d'états possibles (nombre de modes $\times$ 20) que le processus Markov-modulé peut explorer ajoute au calcul de vraisemblance une charge computationnelle supplémentaire.

Notre modèle Markov-modulé a été développé dans le programme d'inférence Bayésienne PhyloBayes-MPI version 1.4 (Lartillot et al. 2013). Afin de réduire efficacement le temps nécessaire à l'échantillonnage des distributions postérieures avec les méthodes MCMC, tout un travail d'optimisation MPI (*Message Passing Interface*) (Snir et al. 1995) a été nécessaire. La figure 10 illustre comment est orchestré le flux d'informations avec un programme MPI. Le processeur maître (M) échantillonne les valeurs de paramètres phylogénétiques (θ) de la distribution postérieure par MCMC et transmet la copie d'une réalisation à chacun des processeurs esclaves (P1 à P6). Ces derniers renvoient par la suite les vraisemblances de θ (L_1 à L_6) conditionnelles aux différents blocs d'alignement qui leur sont respectivement confiés pour que finalement M puisse calculer celle de l'alignement complet ($L(\theta)$).

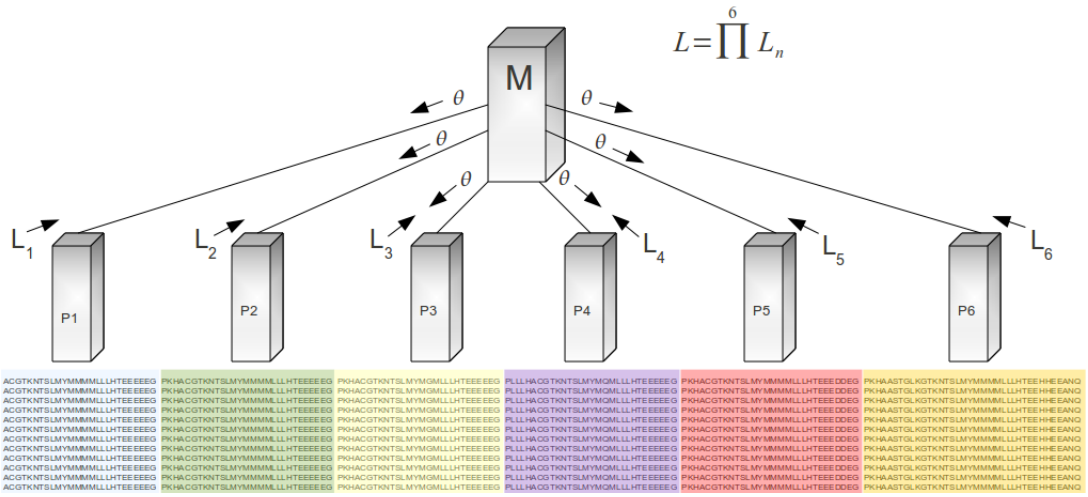


Figure 10 – Illustration simplifiée de la parallélisation du calcul de la vraisemblance (L(θ)) avec les librairies MPI.

Une autre différence fondamentale entre notre modèle et celui de Whelan (2008) porte sur la question de réversibilité en temps (nous reviendrons d'ailleurs sur ce concept dans les matériels et méthodes). Sommairement, une telle supposition a plusieurs avantages sur le plan pratique. Elle permet entre autres d'appliquer le principe de la poulie (Felsenstein 1981) et ainsi de calculer la vraisemblance peu importe la position de la racine. Cependant, elle peut potentiellement amener le processus Markovien continu en temps à transgresser sérieusement le réel processus de substitution biologique. C'est pourquoi nous avons choisi contrairement à Whelan (2008) de ne pas contraindre le noyau stochastique modulé à respecter l'équation du bilan détaillé.

Nous avons finalement utilisé une technique de validation croisée pour comparer différentes configurations de notre modèle Markov-modulé. Ces configurations ont été comparées entre elles et avec des modèles homogènes en temps afin de mesurer leur degré d'accommodation du processus substitutionnel profil-spécifique hétérogène en temps. Nous avons également tenté d'évaluer s'il était possible de simuler l'effet RAS avec un modèle Markov-modulé à la fois pour les changements de mode et pour les changements de vitesse d'évolution.

3 MATÉRIELS ET MÉTHODES

3.1 Modèle Markovien de substitution entre acides aminés

Dans la suite, nous travaillerons exclusivement en acides aminés (le nombre d'états du processus de substitution, noté S, sera donc de 20). Toutefois, une grande partie des développements de modèle présentés ci-dessous, hormis les matrices empiriques de type JTT, sont en principe transposables aux modèles nucléotidiques ($S = 4$).

Afin de modéliser les processus de substitution entre les 20 acides aminés, nous avons intégré à notre modèle le plus général des processus, temps-réversible ; le modèle *General time-reversible* (GTR)(Tavaré 1986). Rappelons que sous cette hypothèse, le taux instantané de substitution d'un état observé a vers un état observé b peut être paramétré comme suit :

$$Q_{ab} = \rho_{ab}\pi_b \qquad a \neq b; \qquad a, b \in [1..S]; \qquad S = 20, \tag{9}$$

où π_b représente la fréquence d'équilibre de b tel que $\sum_{b=1}^{20} \pi_b = 1$ et ρ_{ab} est le taux relatif d'échange entre a et b tel que $\rho_{ab} = \rho_{ba}$, $\forall a > b$. Ces derniers constituent ainsi les entrées d'un vecteur ρ de dimension égale à

$$\frac{S \times (S-1)}{2}. \tag{10}$$

Précisons également ici que l'ensemble des fréquences d'équilibre est habituellement considéré comme étant le profil du processus de substitution.

Tel que déjà introduit, un modèle Markov-modulé assume l'existence de plus d'un processus Markovien de substitution entre acides aminés. Leur profil respectif ainsi que leur nombre espéré *a posteriori* peuvent varier selon le jeu de données. Mais, compte tenu du fait que la charge computationnelle imposée à un modèle Markov-modulé augmente considérablement avec le nombre (G) de processus de substitution, nous avons décidé de restreindre sa dimension à $G = 6$. Nous avons également convenu que l'étude du comportement général du modèle n'exige pas d'estimer les profils directement à partir du jeu de données. Un modèle de mélange (ECG6), entraîné par un algorithme EM développé par Groussin et Lartillot (en préparation), a permis d'estimer 6 matrices empiriques de substitution entre acides aminés sur une base d'alignements de séquences protéiques nucléaires. Puisque la plupart de nos essais ont également été effectués sur des jeux de données nucléaires, nous avons jugé qu'il était approprié d'incorporer à notre modèle les profils et le vecteur de taux relatifs d'échange entre acides aminés caractérisant ces 6 processus de substitution. Les 6 profils sont présentés sous forme de *logos* à la figure 11. Dans ce qui suit, nous ferons référence à l'ensemble des processus de substitution, ou états cachés, correspondants sous l'appellation *HSp6* dans les cas où les 6 profils sont utilisés et sous l'appellation *HSp3* dans les cas ou seulement les profils 1, 5 et 6 sont utilisés.

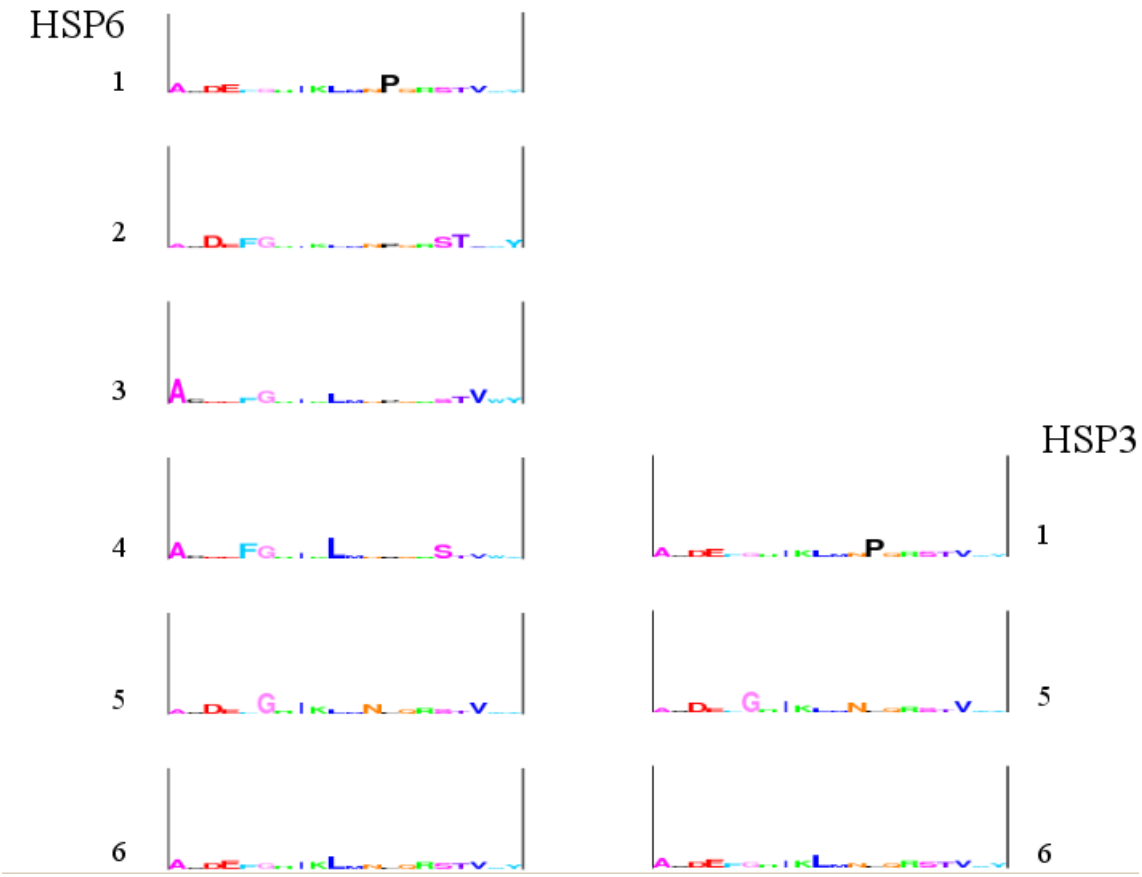


Figure 11 – Représentation sous forme de *logos* des profils des états cachés *HSp6* et *HSp3*. La taille de chacune des lettres est proportionnelle à la fréquence d'équilibre de l'acide aminé correspondant.

3.2 Modèle Markov-modulé des processus de substitution

3.2.1 Généralités sur le modèle

Notre modèle Markov-modulé assume l'existence de différents états cachés, $k = 1 \ldots G$. Chacun étant caractérisé par son propre vecteur de fréquences d'équilibre π^k tel que,

$$\pi^k = (\pi_1^k, \pi_2^k, \ldots, \pi_S^k) \qquad S = 20, \tag{11}$$

$$\sum_{s=1}^{20} \pi_s^k = 1. \tag{12}$$

Ils partagent néanmoins un seul et même vecteur ρ de taux relatifs d'échange symétrique entre acides aminés. Chaque état caché correspond donc à une matrice stochastique M^k, de dimension S x S, définissant un processus spécifique de substitution GTR entre acides aminés. Ainsi pour l'état caché k,

$$\mathrm{M}_{a,b}^k = \rho_{ab}\pi_b^k \qquad a \neq b; \qquad a, b \in [1..S]; \qquad S = 20. \tag{13}$$

Les modulations entre états cachés sont aussi modélisés avec des processus de Markov réversibles en temps sous forme d'une matrice de transition C, de dimension G x G, à l'intérieur de laquelle chaque entrée est définie par un taux relatif d'échange $\eth_{kl}$ et une fréquence d'équilibre η^l. L'équation 13, adaptée au taux de modulation instantané de l'état caché k vers l'état caché l, devient alors

$$C^{k,l} = \eth_{kl}\eta^l \qquad k \neq l; \qquad k, l \in [1..G]. \tag{14}$$

Les fréquences d'équilibre respectives des états cachés 1 à G du vecteur η sont normalisées de manière à respecter la même contrainte que celle imposée aux fréquences d'équilibre des états observés (équation 12) :

$$\sum_{k=1}^{G} \eta^k = 1. \tag{15}$$

Les taux relatifs d'échange entre états cachés sont symétriques, tel que

$$\eth_{kl} = \eth_{lk}, \tag{16}$$

et constituent ainsi les entrées d'un vecteur $\eth$ de dimension égale à

$$\frac{G \times (G-1)}{2}. \tag{17}$$

La combinaison des G matrices $\mathrm{M}^k_{a,b}$ avec la matrice $C^{k,l}$ crée un processus modulé d'ordre SG généralisant tous les taux de transition possibles entre les S états observés et entre les G états cachés. Une matrice Markov-modulée Q constituée par exemple de 2 états cachés, m et n, serait définie par

$$Q = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ S \\ 1 \\ 2 \\ \vdots \\ S \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & \cdots & S & 1 & 2 & \cdots & S \end{array} \\ \begin{pmatrix} - & Q^{m,m}_{1,2} & \cdots & Q^{m,m}_{1,S} & Q^{m,n}_{1,1} & 0 & \cdots & 0 \\ Q^{m,m}_{2,1} & - & \cdots & Q^{m,m}_{2,S} & 0 & Q^{m,n}_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Q^{m,m}_{S,1} & Q^{m,m}_{S,2} & \cdots & - & 0 & 0 & \cdots & Q^{m,n}_{S,S} \\ Q^{n,m}_{1,1} & 0 & \cdots & 0 & - & Q^{n,n}_{1,2} & \cdots & Q^{n,n}_{1,S} \\ 0 & Q^{n,m}_{2,2} & \cdots & 0 & Q^{n,n}_{2,1} & - & \cdots & Q^{n,n}_{2,S} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q^{n,m}_{S,S} & Q^{n,n}_{S,1} & Q^{n,n}_{S,2} & \cdots & - \end{pmatrix} \end{array} \tag{18}$$

et autoriserait ainsi des transitions entre les GS états disponibles.

La matrice Q apparaissant à l'équation 18, généralisée à G états cachés, est définie comme suit :

$$Q_{i,j}^{k,l} = \begin{cases} M_{i,j}^{k} & i \neq j, k = l \\ C^{k,l} & i = j, k \neq l \\ 0 & i \neq j, k \neq l \end{cases}, \tag{19}$$

$$Q_{i,i}^{k,k} = -\left(\sum_{i \neq j} Q_{i,j}^{k,k} + \sum_{k \neq l} Q_{i,i}^{k,l} \right), \tag{20}$$

où $i, j \in [1..S]$ et $k, l \in [1..G]$.

3.2.2 Modulation et réversibilité

Sous les modèles d'évolution moléculaire qui assument que les processus de substitution sont réversibles en temps, la quantité de changements allant de a vers b (trajectoires directes) est égale à la quantité de changements allant de b vers a (trajectoires opposées). Mathématiquement, la réversibilité équivaut à ce que l'on appelle parfois le bilan détaillé : pour un processus Q de fréquence d'équilibre π, alors

$$\pi_a Q_{ab} = \pi_b Q_{ba}. \tag{21}$$

Les modèles de type GTR, comme leur nom l'indique, sont réversibles puisque si l'on paramétrise Q_{ab} en $Q_{ab} = \rho_{ab}\pi_b$, alors

$$\pi_a Q_{ab} - \pi_b Q_{ba} = \pi_a \rho_{ab} \pi_b - \pi_b \rho_{ba} \pi_a \Leftrightarrow \rho_{ab} = \rho_{ba}. \tag{22}$$

Bien que nous ayons présenté, selon les équations 13 et 14, des processus GTR de substitution entre acides aminés et entre états cachés, le résultat de leur combinaison est un modèle modulé non-réversible selon le bilan détaillé suivant :

$$\eta^k \pi_a^k Q_{a,a}^{k,l} \neq \eta^l \pi_a^l Q_{a,a}^{l,k} \qquad \text{car}$$

$$\eta^k \pi_a^k \eth_{kl} \eta^l \neq \eta^l \pi_a^l \eth_{lk} \eta^k,$$

$$\pi_a^k \neq \pi_a^l.$$

La réversibilité est pratique sur le plan technique du fait qu'elle permet d'appliquer le principe de la poulie (Felsenstein 1981) et donc de calculer la vraisemblance peu importe la position de la racine. Elle permet également de structurer la matrice de taux instantané de telle façon qu'elle soit plus facilement diagonalisable (et donc exponentiable par des méthodes numériques). Ce sont les deux raisons principales qui expliquent pourquoi les modèles réversibles sont souvent préférés. Whelan (2008), dont les travaux ont été

introduits précédemment, a d'ailleurs choisi cette option pour implémenter son modèle Markov-modulé en redéfinissant les taux de modulation comme suit :

$$Q_{a,a}^{k,l} = \eth_{kl}\eta^l\pi_a^l. \tag{23}$$

Puisque les fréquences d'équilibre d'un tel modèle sont définies simplement par

$$\zeta_a^k = \eta^k\pi_a^k, \tag{24}$$

le bilan détaillé résultant est alors réversible selon

$$\eta^k\pi_a^k\eth_{kl}\eta^l\pi_a^k = \eta^l\pi_a^l\eth_{lk}\eta^k\pi_a^k \Leftrightarrow \eth_{kl} = \eth_{lk}. \tag{25}$$

Durant ce projet, nous avons implémenté les versions réversibles et non-réversibles du modèle Markov-modulé pour les processus de substitution. Toutefois, nous montrerons l'essentiel des résultats dans le cas non-réversible, qui semble être mieux ajusté aux données observées. Cela dit, il nous a fallu adapter les approches computationnelles nécessaires pour échantillonner de la distribution postérieure avec la méthode MCMC. La section suivante aborde en détail les méthodes numériques sous-jacentes.

3.2.3 Calcul de la vraisemblance sous un modèle Markov-modulé

Certains mouvements de type *pruning* (Felsenstein 1981) de notre algorithme MCMC exigent de calculer la vraisemblance de θ (ensemble des paramètres phylogénétiques du modèle) sachant les données observées D ($p(D \mid \theta)$). Et puisque que les états observés peuvent être dans différents états cachés, la propagation de cette vraisemblance des feuilles de l'arbre jusqu'à la racine doit être adaptée en conséquence. Aussi, dans le cas non-réversible, le calcul du vecteur de fréquences d'équilibre de la matrice Q ainsi que son exponentiation nécessitent des approches différentes de celles utilisées sous la version réversible.

Étant donné un arbre comportant F taxons et un vecteur C_i d'états observés $C_i f_{f\in[1..\mathrm{F}]}$ au site i d'un alignement. La probabilité des données D, ou encore la vraisemblance globale $L(\theta)$, est le produit de la vraisemblance à chacun des sites :

$$L(\theta) = \prod_i p(C_i \mid \theta). \tag{26}$$

Afin de propager la vraisemblance des feuilles de l'arbre jusqu'à la racine, nous avons adapté l'algorithme de *pruning* de manière analogue à celle proposée par Galtier & Jean-Marie (2004). Le calcul est effectué en conditionnant sur chacun des GS états possibles du processus Markov-modulé. Ainsi, si l'on pose X_0 et Z_0 comme étant respectivement les états observés et cachés ancestraux à la racine, la probabilité de C_i est définie comme suit :

$$p(C_i) = \sum_{x\in S}\sum_{z\in\mathrm{G}} p(C_i \mid X_0 = x, Z_0 = z) \times \zeta_x^z, \tag{27}$$

où ζ_x^z, sous le version non-réversible, est calculée par la résolution du système d'équations linéaires avec la méthode d'élimination de Gauss-Jordan (Atkinson 1989)

Suivant la même notation, étant donnée la paire (X,Y) représentant les états d'un noeud ancestral n et $((X_1,Y_1),(X_2,Y_2))$ ceux de ses noeud fils n_1 et n_2 :

$$L_i^n(\theta \mid X = x, Z = z) = \sum_{x_1 \in S} \sum_{z_1 \in G} L_i^{n_1}(\theta \mid X_1 = x_1, Z_1 = z_1) \cdot p(x, z \to x_1, z_1 \mid l_1) \times \sum_{x_2 \in S} \sum_{z_2 \in G} L_i^{n_2}(\theta \mid X_2 = x_2, Z_2 = z_2) \cdot p(x, z \to x_2, z_2 \mid l_2), \quad (28)$$

où l_1 et l_2 sont respectivement les longueurs de branche séparant n de n_1 et de n_2. Les probabilités de transition conditionnelles à l exigent quant à elles de calculer l'exponentielle de la matrice Q. Par exemple, le long de l_1,

$$p(x, z \to x_1, z_1 \mid l_1) = P_{x,x_1}^{z,z_1}(l_1) = [e^{l_1 Q}]_{(x,x_1)(z,z_1)}. \quad (29)$$

Plusieurs méthodes différentes existent pour calculer l'exponentielle d'une matrice (Moler et Van Loan 1978). Sous notre modèle Markov-modulé réversible en temps, la résolution de l'équation 29 peut être effectuée en calculant la forme diagonale de Q. Sous la configuration non-réversible, notre algorithme de diagonalisation par décomposition QR (Wilkinson 1965) est toutefois incompatible. Pour cette raison, nous avons trouvé plus pratique d'utiliser une méthode d'approximation numérique s'appuyant sur l'idée suivante :

$$e^{l_1 Q} = \lim_{y \to \infty} \left(I + \frac{l_1 Q}{2^y} \right)^{2^y}, \quad (30)$$

où I représente la matrice identité. Nous avons évalué expérimentalement que $y = 25$ permet d'obtenir des résultats virtuellement indistinguables de ceux obtenus avec la méthode par diagonalisation.

La toute première étape de notre algorithme de *pruning* Markov-modulé consiste à initialiser aux feuilles f de l'arbre les vraisemblances conditionnelles à l'état observé à chacune des F feuilles avec la valeur 1.0 dans chacun des G états cachés possibles :

$$L_i^f(\theta \mid X_f = x, Z_f = z) = \begin{cases} 1.0 & x = C_i f, \forall z \\ 0 & x \neq C_i f, \forall z \end{cases} \quad f \in [1..\mathrm{F}];\ C_i f, x \in [1..S];\ z \in [1..G]\ . \quad (31)$$

En d'autres termes, cette initialisation revient à sommer les vraisemblances conditionnelles à chacun des états cachés possibles aux feuilles, dans la mesure où ces états cachés sont inconnus.

Rappelons pour terminer cette section que, sous le modèle non-réversible, la valeur de la vraisemblance est dépendante de la trajectoire du processus d'évolution. L'emplacement de la racine devient donc un paramètre important du vecteur θ. Tel que détaillé plus loin, les deux arbres phylogénétiques que nous avons utilisés étaient donc enracinés. Et afin d'avoir la possibilité de calculer efficacement la vraisemblance à partir de n'importe quel noeud, nous avons implémenté dans notre code la méthode proposée par Boussau & Gouy (2006) pour les processus non-réversibles en temps.

3.2.4 Modulation et variation de la vitesse d'évolution entre positions

Il est aujourd'hui bien connu que les modèles accommodant l'hétérogénéité des vitesses de substitution à travers les sites offrent un meilleur ajustement avec les données observées. Le modèles RAS introduit par Yang (1993), suggérant que ces vitesses entre sites sont tirées d'une distribution gamma, est d'ailleurs la plupart du temps intégré d'emblée dans les modèles probabilistes. En fait, du point de vue pratique, il est plus commode d'approximer la distribution gamma continue en la discrétisant en m catégories ($m = 4$ est souvent une approximation appropriée). Mais puisque que l'affiliation respective des sites à chacune des catégories est un mécanisme inconnu, l'équation 27 est adaptée de façon à ce que la vraisemblance soit sommée sur les m catégories de la gamma discrétisée (D-Gam) comme suit :

$$p(C_i \mid \theta) \simeq \sum_{m=1}^{4} w_m \, p(C_i \mid \theta, r = \overline{r_m}), \tag{32}$$

où $\overline{r_m}$ est le taux median et w_m est le poids attribué à la mth catégorie de la D-Gam. Dans notre cas, $w_m = 1/4 \quad \forall m$.

Mathématiquement, la vitesse d'évolution au site i (r_i) agit comme un multiplicateur de longueur de branche (l) pour calculer la probabilité de transition après un temps t selon

$$P(t) = e^{r_i l Q}. \tag{33}$$

Mais dans le cas de notre modèle Markov-modulé pour les processus de substitution, cette vitesse r_i tirée de la D-Gam agit non seulement sur la vitesse des substitutions entre acides aminés mais également sur celle des modulations entre états cachés. Cette modélisation simultanée des deux types de vitesse d'évolution n'est pas nécessairement en accord avec la réalité biologique. Une alternative est d'appliquer la D-Gam uniquement pour les transitions entre acides aminés et d'assumer que la vitesse de modulation entre états cachés est la même à travers tous les sites. C'est l'option choisie par Huelsenbeck (2002) avec son modèle covarion. Celui-ci accommode la coexistence de la D-Gam avec le phénomène de transitions entre états cachés en élaborant une matrice modulante Q différente pour chaque catégorie de vitesse de substitution entre états observés. La vitesse de modulation entre les états cachés est quant à elle indépendante de la catégorie et est la même pour tous les sites. Cela dit, cette alternative ainsi que la notre sont toutes deux défendables.

Finalement, simplement rappeler ici que sous notre modèle, les sites dits rapides le sont à la fois en terme de substitutions et en terme de transitions entre états cachés. Et que le même raisonnement s'applique pour les sites lents.

3.2.5 Modèle Markov-modulé covarion

Rappelons que les modèles de type covarion développés par Tuffley & Steel (1998), Huelsenbeck (2002), Galtier (2001) et Wang et al. (2007) ne tentent que d'accommoder

les variations de taux d'évolution à travers les sites et au cours du temps. Notre modèle Markov-modulé quant à lui n'accommode pas la dimension temporelle des vitesses d'évolution. Il autorise un site à moduler entre différents processus de substitution profil-spécifiques mais toujours sous le même taux de substitution site-spécifique.

Nous proposons donc ici un deuxième modèle Markov-modulé pour les processus de substitution, sans la D-Gam mais qui intègre une dimension covarion de type ON (vitesse d'évolution non-nulle) et OFF (vitesse d'évolution nulle) comme celle proposée par Tuffley & Steel (1998). En plus des états cachés caractérisés par des profils différents, nous avons combinés à la matrices Q un état caché supplémentaire (nommé 0) à l'intérieur duquel la vitesse d'évolution est nulle. La matrice modulante résultante, que nous avons nommée Q^{cov}, est définie comme suit :

$$Q_{i,j}^{cov(k,l)} = \begin{cases} M_{i,j}^{k} & i \neq j, k = l \neq 0 \\ 0 & i \neq j, k = l = 0 \\ C^{k,l} & i = j, k \neq l \\ 0 & i \neq j, k \neq l \end{cases}, \tag{34}$$

$$Q_{i,i}^{cov(k,k)} = -\left(\sum_{i \neq j} Q_{i,j}^{cov(k,k)} + \sum_{k \neq l} Q_{i,i}^{cov(k,l)} \right). \tag{35}$$

La valeur d'ajustement aux données observées de Q comparée celle de Q^{cov} permet entre autres d'évaluer partiellement si accommoder l'hétérogénéité des taux d'évolution à travers le temps permet de compenser la perte de la D-Gam. L'implémentation de ces deux configurations permet également d'évaluer laquelle est la mieux adaptée pour détecter des événements de divergences fonctionnelles.

3.2.6 Modèles testés et terminologies

Durant le présent projet, nous avons évalué le comportement de différentes configurations de notre modèle Markov-modulé et avons comparé leur performance respective à celles d'autres modèles déjà existants. Dans les sections qui suivent, nous ferons référence à ces différentes configurations en respectant leur notation respective apparaissant dans les tableaux 1 à 3. Ces tableaux résument les caractéristiques respectives de chacun des 9 modèles relatives à leur degré d'hétérogénéité et à leur réversibilité.

Tableau 1 – Modèles avec un seul processus de substitution

	LG	LG+Γ	GTR+Γ
$(\pi + \rho)\, fixes$[a]	✓	✓	-
$(\pi + \rho)\, libres$	-	-	✓
D-Gam	-	✓	✓
Réversible	✓	✓	✓

[a] π et ρ estimés par Le & Gascuel (2008).

Tableau 2 – Modèles CAT[a]

	CAT3	CAT6	CAT3+Γ	CAT6+Γ	CAT3-GTR+Γ	CAT6-GTR+Γ
États cachés $HSp6$ avec ρ fixe	-	✓	-	✓	-	-
États cachés $HSp3$ avec ρ fixe	✓	-	✓	-	-	-
États cachés $HSp6$ avec ρ libre	-	-	-	-	-	✓
États cachés $HSp3$ avec ρ libre	-	-	-	-	✓	-
D-Gam	-	-	✓	✓	✓	✓
Réversible	✓	✓	✓	✓	✓	✓

[a] Lartillot & Philippe (2004).

Tableau 3 – Modèles Markov-modulés

	$MM3_\Gamma$	$MM6_\Gamma$	$MM3^r_\Gamma$	$MM6^r_\Gamma$	$MM3_{cov}$	$MM6_{cov}$
États cachés $HSp6$ avec ρ fixe	-	✓	-	✓	-	-
États cachés $HSp3$ avec ρ fixe	✓	-	✓	-	-	-
D-Gam	✓	✓	✓	✓	-	-
Réversible	-	-	✓	✓	-	-
covarion	-	-	-	-	✓	✓

3.3 Priors sur les paramètres du vecteur θ

- Chaque longueur l de branche j est tirée *a priori* d'une distribution gamma γ paramétrée avec le paramètre de forme $\nu > 0$ et le paramètre d'échelle $\mu > 0$. La densité *a priori* sur l'ensemble des branches, notée l, est donc

$$p(\mathbf{l}) = \prod_j \gamma_{\nu,\mu}(l_j) = \prod_j \frac{\mu^\nu}{\Gamma(\nu)} l_j^{\nu-1} e^{-\mu l_j}, \tag{36}$$

  où $\Gamma(\nu)$ est la fonction gamma évaluée à ν. Une distribution exponentielle de moyenne 1 a été utilisée pour les paramètres ν et μ :

$$p(\nu) = e^{-\nu}, \tag{37}$$

$$p(\mu) = e^{-\mu}. \tag{38}$$

- La distribution gamma (D-Gam discrétisée en 4 catégories) utilisée pour modéliser la distribution des vitesses d'évolution à travers les sites est modulée par son paramètre de forme, α, également tiré *a priori* d'une distribution exponentielle de moyenne 1 :

$$p(\alpha) = e^{-\alpha}. \tag{39}$$

- Lorsqu'ils sont laissés libres, les taux relatifs d'échange entre états observés (ρ_{ab}) sont uniformément distribués *a priori*. Alternativement, ils sont fixés aux valeurs telles que spécifiées par le modèle empirique LG (Quang et al. 2008c, Quang et al. 2008d) ou par le modèle empirique de mélange ECG6 (Groussin et Lartillot, in prep). Dans les deux cas, ils sont normalisés et ensuite contraint à ce que leur somme soit telle que

$$\sum_{1 \le a < b \le S} \rho_{ab} = \frac{S(S-1)}{2}. \tag{40}$$

- Étant donné que nous n'avions aucune information *a priori* sur les taux relatifs d'échange entre états cachés, nous avons proposé que la prior sur les $\eth_{kl}$ soit choisie de manière à ce que les taux de transitions entre états observés soient en moyenne les mêmes que ceux entre les états cachés. Nous avons donc posé, sous l'hypothèse d'uniformité des vecteurs ρ, $\eth$, π^k et η, et d'un processus non-réversible en temps,

$$Q^{k,k}_{i,j} = Q^{k,l}_{i,i} \quad i \neq j;\ k \neq l \qquad \Rightarrow \rho_{ij}\pi^k_j = \eth_{kl}\eta^l,$$

  et

$$\pi^k_j S = \eta^l G = 1.$$

  Et donc, puisque

$$\eth_{kl} = \frac{\rho_{ij}\pi^k_j}{\eta^l} = \frac{\rho_{ij} G}{S},$$

  et que

$$\rho_{ij} = 1,$$

$$\Rightarrow \eth_{kl} = \frac{G}{S}.$$

  La même logique appliquée au processus réversible génère la relation

$$\eth_{kl} = G.$$

Sous les modèles Markov-modulés non-réversibles en temps, les valeurs de $\eth_{kl}$ indépendantes et identiquement distribuées (i.i.d), sont donc tirées d'une distribution exponentielle de moyenne G/S :

$$\eth_{kl} \sim Exp(S/G) \qquad S = 20, \tag{41}$$

tandis que sous le modèle réversible,

$$\eth_{kl} \sim Exp(1/G) \qquad S = 20. \tag{42}$$

- Le vecteur de fréquences d'équilibre des états cachés (η) a comme prior une distribution de Dirichlet uniforme non-informative :

$$\eta \sim Dirichlet(1, 1, \ldots, 1). \tag{43}$$

- Sous le modèle GTR avec un seul processus de substitution, le profil (libre) est également tiré d'une distribution de Dirichlet :

$$\pi^k \sim Dirichlet(1, 1, \ldots, 1) \quad G = 1. \tag{44}$$

Pour tous les autres modèles dont il sera question dans ce travail, les profils des différents processus de substitution sont fixes et selon le cas ont été estimés empiriquement sous le modèle LG ou sous le modèle de mélange ECG6 de Groussin et Lartillot (in prep).

3.4 Approches computationnelles et MCMC

3.4.1 Parallélisation

Notre programme a été implémentée en C++ à l'intérieur du programme d'inférence Bayésienne PhyloBayes-MPI version 1.4 (Lartillot et al. 2013). Sans les avantages de la technologie MPI, notre modèle Markov-modulé Bayésien n'aurait probablement jamais vu le jour. Avec un jeu de données protéique ($S = 20$) et par exemple 6 états cachés ($G = 6$), le nombre d'états ancestraux possibles passe de 20 à 120. Ainsi, à chaque site, l'algorithme de *pruning* doit prendre en charge ces 120 états. Ce qui, du point de vue computationnel, serait évidemment un désavantage majeur avec un seul processeur.

L'algorithme de *pruning* (équation 27) est un calcul récursif de vraisemblance partant des feuilles de l'arbre jusqu'à la racine. Et puisque qu'à chacun des noeuds les valeurs de vraisemblance sont conditionnelles à chacun des GS états possibles du processus modulé, la complexité en temps de cet algorithme est $\mathcal{O}((GS)^2)$. Ce qui signifie que le temps qu'il demande pour compléter le calcul croît de manière quadratique avec l'augmentation du nombre d'états. Le calcul de la probabilité de transition entre deux états après un temps t défini à l'équation 29 (pour le processus non-réversible en temps) nécessite également un temps de résolution non négligeable ; une multiplication de matrice par elle-même, élevée à la puissance 25. C'est une complexité algorithmique de $\mathcal{O}((GS)^3)$.

Tel que déjà expliqué sommairement plus haut, la parallélisation MPI consiste à répartir, sous le contrôle d'un processeur maître, à travers un ensemble d'esclaves, les calculs de $L(\theta)$ ainsi que l'échantillonnage d'histoires substitutionnelles (*mapping*) site-spécifiques de la distribution postérieure conditionnelle. Chacun des esclaves renvoie au maître la vraisemblance de θ sachant le bloc d'alignement observé sous leur charge respective ainsi que des statistiques suffisantes (données augmentées) (Rodrigue et al. 2008) calculées à partir de leurs *mapping*. Le maître quant à lui est responsable de sommer les log-vraisemblances et les statistiques suffisantes transmis par les esclaves et d'échantillonner les θ de la distribution postérieure.

Nos calculs ont été exécutés sur les serveurs (Cottos/Université de Montréal (UdeM), Briaré/UdeM et Mammouth (mp2)/Université de Sherbrooke) du Réseau québécois de calcul de haute performance (RQCHP) et sur un serveur privé à l'UdeM (sirocco). Nous avons effectué une série d'essais visant à déterminer quel était le degré de parallélisation optimal dans le contexte de nos analyses. Il s'est avéré que la configuration la plus performante est celle utilisant 3 noeuds avec chacun 24 coeurs (pour un total de 72 coeurs) sur le serveur mp2. Globalement, avec un jeu de données de 25 640 positions comportant chacune 22 acides aminés alignés, le temps moyen nécessaire pour effectuer un seul cycle MCMC sous $\mathrm{MM6}_{\Gamma}$ avec ce degré de parallélisation est de 28 minutes. Donc, après le temps maximal alloué sur ce serveur, soit 120 heures, seulement 257 points échantillonnés ont été sauvegardés. D'où la nécessité de repartir les chaînes à chaque cycle de 5 jours pour obtenir un l'échantillon final suffisamment représentatif de la distribution postérieure.

3.4.2 Échantillonnage de la distribution postérieure

Le théorème de Bayes (équation 1) stipule que les paramètres du modèle (collectivement désignés par θ) sont distribués *a posteriori* selon le produit de leur vraisemblance (section 3.2.3) par leur prior. L'objectif du MCMC en phylogénie moléculaire est d'obtenir un échantillon de cette distribution postérieure conditionnelle aux données empiriques. Dans notre cas, nous avons procédé par alternance entre deux méthodes MCMC. La première basée sur l'algorithme de Métropolis-Hastings (MH) et la seconde sur la procédure de *Gibbs sampling*. Nous avons discuté brièvement du principe d'échantillonnage par MCMC dans l'introduction. Mais son application aux méthodes de reconstruction phylogénétique est davantage détaillée dans Huelsenbeck et al. (2002), Rannala (2002), Yang & Rannala (1997), Larget & Simon (1999), Mau (1996) et Li (1996).

Précisons ici concernant la méthode MH que nous avons utilisé deux approches différentes pour calculer les vraisemblances de l'état proposé (θ^*) et de l'état initial (θ) du ratio de Métropolis. En résumé, notre programme, en mode marginal, intègre la vraisemblance sur toutes les histoires substitutionnelles possibles avec l'algorithme de *pruning* décrit à la section 3.2.3. Alternativement, en mode augmenté, une histoire substitutionnelle détaillée est premièrement tirée de la distribution *a posteriori* suivant l'algorithme décrit par Rodrigue et al. (2008). Puis, les paramètres du modèle sont rééchantillonnés conditionellement cette histoire substitutionnelle. L'avantage de procéder ainsi est que la vraisemblance augmentée (la probabilité de l'histoire substitutionnelle détaillée conditionellement aux valeurs de paramètres) se calcule très rapidement comparativement à l'algorithme de *pruning*. Étant donné que les statistiques suffisantes calculées à partir des histoires substitutionnelles (essentiellement, les nombres totaux de transitions entre chaque paire d'états, les temps d'attente dans chaque état du processus Markov-modulé et la fréquence de ces états à la racine) sont très compactes. Leur probabilité sachant θ est effectivement obtenue avec une série d'opérations mathématiques relativement simplistes. L'inconvénient avec le mode augmenté est qu'il peut y avoir des dépendances fortes entre histoire substitutionnelle et valeurs de paramètres. C'est un point d'ailleurs sur lequel nous reviendrons et aborderons plus en détail dans la discussion.

À chacun des cycles du MCMC, différentes valeurs sont proposées pour les différents paramètres du vecteur θ. Les mouvements proposés sont les suivants :

- Deux approches distinctes sont utilisées en alternance pour bouger les longueurs de branches. La première consiste à bouger celles-ci consécutivement par mouvements de type multiplicatif tel que déjà décrit dans Lartillot (2006). C'est l'approche que nous avons utilisée pour un échantillonnage MCMC de type MH en mode marginal. La seconde exige premièrement de tirer une histoire substitutionnelle Ξ de la distribution postérieure conditionnelle $p(\Xi \mid D, \theta)$ avant de rééchantillonner les longueurs de branches conditionnellement à Ξ et les autres valeurs de paramètres du vecteur θ. Cette seconde approche constitue en fait le module d'un échantillonnage de Gibbs conjugué détaillé dans Lartillot (2006). La première étape du module étant celle de l'augmentation des données et la deuxième celle de l'échantillonnage Gibbs.

- L'échantillonnage des paramètres μ, α, ρ_{ab} et $\eth_{kl}$ est accompli par MH en mode augmenté avec des mouvements de type multiplicatif.
- Les mouvements sur les vecteurs de type profil, π^k et η, sont contraints à ce que la somme de leurs entrées soit égale à 1.0. Nous avons par conséquent utilisé une procédure adaptée pour ce type de paramètre consistant principalement à coupler les mouvements sur les entrées (Lartillot 2006). Elle consiste à choisir en premier lieu une ou des paires d'entrées du vecteur. Ensuite, si une des valeurs de la paire est bougée de $+x$ alors la seconde valeur couplée est bougée de $-x$.

3.4.3 Réglage du MCMC

Un cycle de MCMC est défini comme étant une série de mouvements sur les différentes valeurs de paramètres d'un modèle. Les mouvements sont réglés de manière à ce que le taux d'acceptation soit entre 25% et 30%. Et afin d'accélérer l'atteinte de la convergence des chaînes et de réduire leur temps de décorrélation respectif, certains paramètres sont soumis à des mouvements réglés avec deux niveaux d'amplitude ; l'un de faible amplitude avec un taux d'acceptation de 60-70% et l'autre d'amplitude plus élevée avec un taux d'acceptation de 5-15%. Les chaînes MCMC sont monitorées en sauvegardant à chaque cycle (dépendamment du modèle) les valeurs de log-vraisemblance, $\bar{\eth}_{kl}$, $\bar{\rho_{ab}}$, $\sigma^2_{\eth_{kl}}$, $\sigma^2_{\rho_{ab}}$, $\sigma^2_{\eta^k}$, α, et la longueur de l'arbre ($\mathbf{l}$) dans un fichier trace. Pour chacun des modèles testés, deux chaînes sont lancées en parallèle afin d'évaluer leur convergence par visualisation de leur trace avec le programme *Gnuplot* (http ://www.gnuplot.info).

Les échantillons de la distribution postérieure sont sous-échantillonés afin de procéder à diverses analyses Bayésiennes d'intérêt. Les conditions de sous-échantillonnage (s.e.) pour chacun des modèles testés sont présentés à l'intérieur des tableaux 4 à 6. Les valeurs moyennes postérieures des différents paramètres sont très révélatrices du comportement d'un modèle. Dans notre cas, celles sur lesquelles nous nous sommes penchées plus particulièrement sont les moyennes postérieures sur les $\eth_{kl}$ et les η^k. Pour visualiser et comparer celles-ci, nous avons développé en LaTeX (http ://www.latex-project.org/) un programme *Bubble plot* et adapté en PostScript l'outil graphique *Sequence logos* (Schneider & Stephens (1990)).

Les points mis de coté avant le début d'un s.e. constituent la phase dite de *burn-in*. Elle correspond à l'ensemble de points non représentatifs de la distribution postérieure ciblée (autrement dit, avant que la chaîne n'ait atteint sa phase dite stationnaire). La fréquence des s.e. est intimement liée au temps de décorrélation entre deux points. Il existe des approches computationnelles efficaces pour estimer ce temps (Gilks & Roberts 1996) mais pour les besoins du présent projet, nous avons évalué qu'un s.e. à chaque 10 cycles permettant d'obtenir un échantillon final de 100 points était approprié. Pour visualiser les distributions postérieures des proportions espérées de transitions entre états cachés par rapport au nombre total de transitions (i.e. prédites par Q ; consulter la section 3.7 pour la définition de $P(Q)$), nous avons construit des histogrammes avec le programme R (http ://www.r-project.org/). Et puisque qu'une taille importante d'échantillon est nécessaire pour obtenir

un histogramme significatif, nous avons " brulé " moins de points (tout en demeurant dans la zone de convergence) et augmenté la fréquence des s.e. à 1.

Tableau 4 – Conditions de s.e. post-MCMC - Microsporidies - Modèles avec un seul processus de substitution

	LG	LG+Γ	GTR+Γ
Longueur des chaînes MCMC[a]	1 500	1 500	1 500
Burn-in[a]	500	500	500
Fréquence des sous-échantillonnages[b]	10	10	10

[a]En terme de cycles.
[b]Nombre de cycles entre deux sous-échantillonnages à partir de la fin du *burn-in* jusqu'à la fin de la chaîne.

Tableau 5 – Conditions de s.e. post-MCMC - Microsporidies - *HSp6*

	CAT6	CAT6+Γ	CAT6-GTR+Γ	$MM6_{\Gamma}$	$MM6^{r}_{\Gamma}$	$MM6_{cov}$
Longueur des chaînes MCMC[a]	1 500	1 500	1 500	3 500	3 500	2 100
Burn-in 1[b]	-	-	-	500	-	1000
Burn-in 2[c]	500	500	500	2 500	2 500	1 600
Fréquence des sous-échantillonnages[d]	10	10	10	10	10	5

[a]En terme de cycles.
[b]Nombre de cycles éliminés pour représenter la distribution postérieure des $P(Q)$.
[c]Nombre de cycles éliminés pour, estimer les moyennes postérieures des paramètres, procécer aux tests prédictifs *a posteriori* et comparer les modèles.
[d]Nombre de cycles entre deux sous-échantillonnages à partir de la fin du *burn-in* jusqu'à la fin de la chaîne.

Tableau 6 – Conditions de s.e. post-MCMC - Microsporidies - *HSp3*

	CAT3	CAT3+Γ	CAT3-GTR+Γ	$MM3_{\Gamma}$	$MM3^{r}_{\Gamma}$	$MM3_{cov}$
Longueur des chaînes MCMC[a]	1 500	1 500	1 500	3 500	3 500	6 100
Burn-in 1[b]	-	-	-	500	-	3000
Burn-in 2[c]	500	500	500	2 500	2 500	4 500
Fréquence des sous-échantillonnages[d]	10	10	10	10	10	20

[a]En terme de cycles.
[b]Nombre de cycles éliminés pour représenter la distribution postérieure des $P(Q)$.
[c]Nombre de cycles éliminés pour, estimer les moyennes postérieures des paramètres, procécer aux tests prédictifs *a posteriori* et comparer les modèles.
[d]Nombre de cycles entre deux sous-échantillonnages à partir de la fin du *burn-in* jusqu'à la fin de la chaîne.

Tableau 7 – Conditions de s.e. post-MCMC - *Mesostigma - HSp6*

	$MM6_\Gamma$
Longueur des chaînes MCMC[a]	2 500
Burn-in 1[b]	1 400
Burn-in 2[c]	1 500
Frequence des sous-échantillonnages[d]	10

[a]En terme de cycles.
[b]Nombre de cycles éliminés pour représenter la distribution postérieure des $P(Q)$.
[c]Nombre de cycles éliminés pour, estimer les moyennes postérieures des paramètres, procécer aux tests prédictifs *a posteriori* et comparer les modèles.
[d]Nombre de cycles entre deux sous-échantillonnages à partir de la fin du *burn-in* jusqu'à la fin de la chaîne.

3.5 Calcul de la log vraisemblance pseudomarginale · LPML

Nous avons utilisé une méthode de validation croisée pour mesurer et comparer l'ajustement des différents modèles testés au jeu de données de microsporidies. Cette méthode, connue sous le nom de *Conditional Predictive Ordinate* (CPO)(Gelfand 1995 ; Lewis et al. 2013), est très bien adaptée pour sélectionner des modèles conçus dans un contexte Bayésien. Par une approche prédictive *a posteriori*, elle évalue l'ajustement d'un modèle M aux données C_i d'un site i conditionnellement aux données observées aux niveaux de tous les autres sites (C_{-i}). La valeur de CPO à un site i est ainsi définie par

$$\mathrm{CPO}_i = p(C_i|C_{-i}, M) = \int p(C_i \mid \theta, M)p(\theta \mid C_{-i}, M)d\theta, \tag{45}$$

et peut être estimée par s.e. d'un chaîne MCMC. Étant donné un sous-échantillon de taille A,

$$\widetilde{\mathrm{CPO}_i} = \left(\frac{1}{A}\sum_{a=1}^{A}\frac{1}{p(C_i \mid \theta^a)}\right)^{-1}. \tag{46}$$

Mais, puisque la quantité $p(C_i \mid \theta^a)$ n'est disponible qu'en échelle logarithmique, il est plus pratique d'évaluer l'équation 46 comme suit :

$$\log\widetilde{\mathrm{CPO}_i} = \log A \ + \log p(C_i \mid \theta^a)_{min} \ - \log\sum_{a=1}^{A} e^{\log p(C_i|\theta^a)_{min} - \log p(C_i|\theta^a)} \tag{47}$$

où $\log p(C_i \mid \theta^a)_{min} = min\{\log p(C_i \mid \theta^a)\}_{a\in[1..A]}$. La somme des $\log \widetilde{\mathrm{CPO}_i}$ à travers tous les sites permet d'obtenir la valeur

$$\mathrm{LPML} = \sum_i \log \widetilde{\mathrm{CPO}_i}, \tag{48}$$

qui est la mesure globale utilisée pour comparer les différents modèles. Et pour connaître δ_{CPO_i}, le gain moyen en CPO par site d'un modèle $\mathcal{M}_1$ par rapport à un modèle $\mathcal{M}_2$, nous appliquons la formule suivante :

$$\delta_{CPO_i} = e^{\left(\frac{\mathrm{LPML}_{\mathcal{M}_1} - \mathrm{LPML}_{\mathcal{M}_2}}{N}\right)} \tag{49}$$

où N représente le nombre de sites.

3.6 Tests prédictifs *a posteriori*

Dans le but de saisir des signaux de divergences fonctionnelles sur des branches spécifiques de la lignée des microsporidies, nous avons opté pour une approche prédictive *a posteriori* similaire à celle employée par Blanquart & Lartillot (2008). Les tests statistiques que nous avons utilisés exigent premièrement l'obtention d'échantillons d'histoires substitutionnelles Ξ tirées de la distribution postérieure observée (contrainte aux données observées) où $\Xi^{\mathrm{o}} \sim p(\Xi^{\mathrm{o}} \mid \theta, D)$ et de la distribution prédictive *a posteriori* (non contrainte) où $\Xi^{\mathrm{s}} \sim p(\Xi^{\mathrm{s}} \mid \theta)$. La statistique calculée à partir de Ξ que nous avons jugée comme étant potentiellement révélatrice d'un signal de divergences fonctionnelles sur une branche donnée j à travers les n sites est la proportion P_j de transitions entre états cachés par rapport au nombre total de transitions. Cette proportion étant définie par le rapport entre le nombre de transitions entre états cachés (H_j) et le nombre total de transitions (nombre de transitions entre états cachés + nombre de transitions entre états observés (O_j)) :

$$P_j = \frac{\sum_i^n H_{ij} + 1}{\sum_{i=1}^n H_{ij} + \sum_{i=1}^n O_{ij} + 2}. \tag{50}$$

$$O_j = \sum_i^n O_{ij} \tag{51}$$

et

$$H_j = \sum_i^n H_{ij}. \tag{52}$$

La proportion **globale** de transitions entre états cachés pour l'ensemble des J branches,

$$P = \frac{\sum_j^J H_j + 1}{\sum_j^J H_j + \sum_j^J O_j + 2} \tag{53}$$

permet ainsi de calculer les déviations standards sur les P_j,

$$\sigma_j = \sqrt{\frac{P * (1 - P)}{H_j + O_j + 1}}, \tag{54}$$

et d'introduire une statistique normalisée,

$$Z_j = \frac{P_j - P}{\sigma_j}. \tag{55}$$

Étant donné que notre analyse prédictive *a posteriori* s'intérese principalement aux modulations entre des profils de préférences en acides aminés, nous avons appliqué un filtre sur les modulations de type covarion sous les modèles $MM3_{cov}$ et $MM6_{cov}$. La figure 12 illustre le principe de filtration que nous avons utilisé. L'état caché à l'intérieur duquel le taux d'évolution est nul est représenté par **0**. Les modulations covarion filtrées sont celles caractérisées par des aller-retour avec celui-ci tel qu'illustré dans le haut de la figure 12. Si **0** sert d'"escale" entre deux états cachés différents, alors nous considérons la séquence de transitions comme **une seule** modulation.

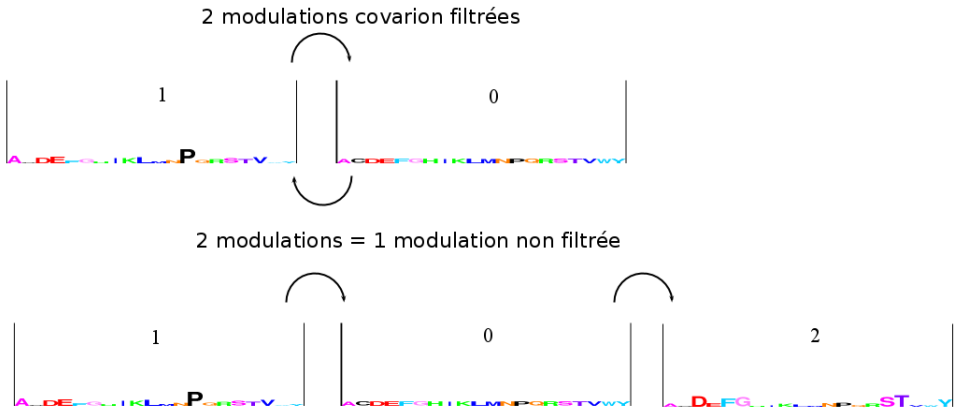


Figure 12 – Application d'un filtre sur les modulations covarion sous le modèle MM_{cov}. **0** représente l'état caché OFF.

Le premier test prédictif *a posteriori* que nous avons mis au point consiste à calculer Z_j^{o} et Z_j^{s} pour chacun des points du s.e.. Afin d'évaluer si les Z_j^{o} d'une branche donnée sont significativement différents de ceux tirés de la distribution nulle (l'hypothèse de nullité), nous avons estimé une valeur p prédictive *a posteriori* en comptant le nombre de points où $Z_j^{\text{s}} \geq Z_j^{\text{o}}$. Ces valeurs p, selon leur valeur de significativité, sont directement mises en évidence sur des arbres colorés construits avec un programme combiné C++/LaTeX. Les longueurs de branches de ces arbres sont des estimés de leur moyenne *a posteriori* respective.

La seconde statistique potentiellement révélatrice du taux élévé de divergences fonctionnelles chez les microsporidies cible le signal fossilisé directement dans les données observées. L'idée est de calculer, pour chacun des n sites i, la proportion $\wp_i$ d'états observés dans le groupe des microsporodies qui sont dans un état caché différent de celui des

autres taxons. Moyennées sur tous les sites, nous obtenons la statistique suivante :

$$\mathcal{F} = \frac{1}{n} \sum_{i=1}^{n} \wp_i \tag{56}$$

où $\wp_i = 0$ si les états observés dans les taxons extérieurs au groupe des microsporidies ne sont pas tous dans le même état caché. Similairement à la statistique Z_j, une valeur p prédictive *a posteriori* est estimée en comptant le nombre de points où $\mathcal{F}^s \geq \mathcal{F}^o$. L'hypothèse de nullité est rejetée si valeur $p < 0.05$.

3.7 Vérification de la corrélation entre Q et Ξ

Puisque chacune des substitutions et modulations partant de la racine jusqu'aux feuilles de l'arbre dépend du générateur stochastique Q, nous avons jugé bon d'intégrer à notre plan de contrôle de la qualité une vérification de la corrélation entre la proportion de modulations entre états cachés prédites par Q ($P_j(Q)$) et celle calculée à partir des histoires substitutionnelles tirées de la distribution postérieure ($P_j(\Xi)$). En théorie, l'on doit s'attendre à ce que ces deux valeurs soient rapprochées.

Afin d'expliquer la procédure à suivre pour calculer $P_j(Q)$, considérons un seul point a d'un sous-échantillon postérieur de taille A. Posons Q_o^a, le taux total de transition entre états observés défini par

$$Q_o^a = \sum_{k}^{G} \sum_{i}^{S} \sum_{j}^{S} Q_{i,j}^{(k,k)^a} \quad i \neq j, \tag{57}$$

et Q_h^a, le taux total de transition entre états cachés tel que

$$Q_h^a = \sum_{k}^{G} \sum_{l}^{G} \sum_{i}^{S} Q_{i,i}^{(k,l)^a} \quad k \neq l. \tag{58}$$

Ainsi pour ce point a, nous avons

$$P_j(Q)^a = \frac{Q_h^a}{Q_o^a + Q_h^a}, \tag{59}$$

et avec l'ensemble des A points, nous pouvons obtenir un estimé de la moyenne postérieure de cette proportion :

$$P_j(Q) = \frac{1}{A} \sum_{a=1}^{A} P_j(Q)^a. \tag{60}$$

Afin de calculer $P_j(\Xi)$ pour une branche donnée, nous calculons tout d'abord, pour chacun des points a, $P_j(\Xi)^a$ selon l'équation 50 et moyennons ensuite sur les A points :

$$P_j(\Xi) = \frac{1}{A} \sum_{a=1}^{A} P_j(\Xi)^a. \tag{61}$$

Et finalement, moyenné sur l'ensemble des J branches, le résultat de

$$P_j(\Xi)\,\text{moyen} = \frac{1}{J}\sum_{j=1}^{J} P_j(\Xi) \tag{62}$$

est comparé avec $P_j(Q)$.

3.8 Arbres phylogénétiques et données

Les résultats présentés à l'intérieur du présent mémoire ont été obtenus à partir des deux jeux de données protéiques suivants :

Microsporidies

Ce jeu de données constitué de 25 640 positions alignées est celui sur lequel nous avons réalisé la plupart de nos essais. Il résulte de la concaténation de protéines nucléaires conservées échantillonnées chez 22 espèces d'eucaryotes ; 6 microporidies (*Antonospora locustae*, *Nematocida parisii*, *Enterocytozoon bieneusi*, *Nosema ceranae*, *Encephalitozoon intestinalis*, *Encephalitozoon cuniculi*), 5 ascomycètes (*Aspergillus nidulans*, *Neurospora crassa*, *Candida albicans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces*), 2 basidiomycètes (*Laccaria bicolor* et *Puccinia graminis*), 3 zygomycètes (*Mucor circinelloides*, *Rhizopus oryzae*, *Phycomyces blakesleeanus*), 2 chytrides (*Batrachochytrium dendrobatidis* et *Spizellomyces punctatus*) et 4 autres organismes eucaryotes (*Nematostella vectens*, *Amphimedon queenslandica*, *Monosiga brevicollis* et *Capsaspora owczarzaki*) constituant le *outgroup*. L'arbre phylogénétique enraciné correspondant (figure 13) ainsi que l'alignement nous ont été aimablement fournis par Capella-Gutirrez et al. (2012).

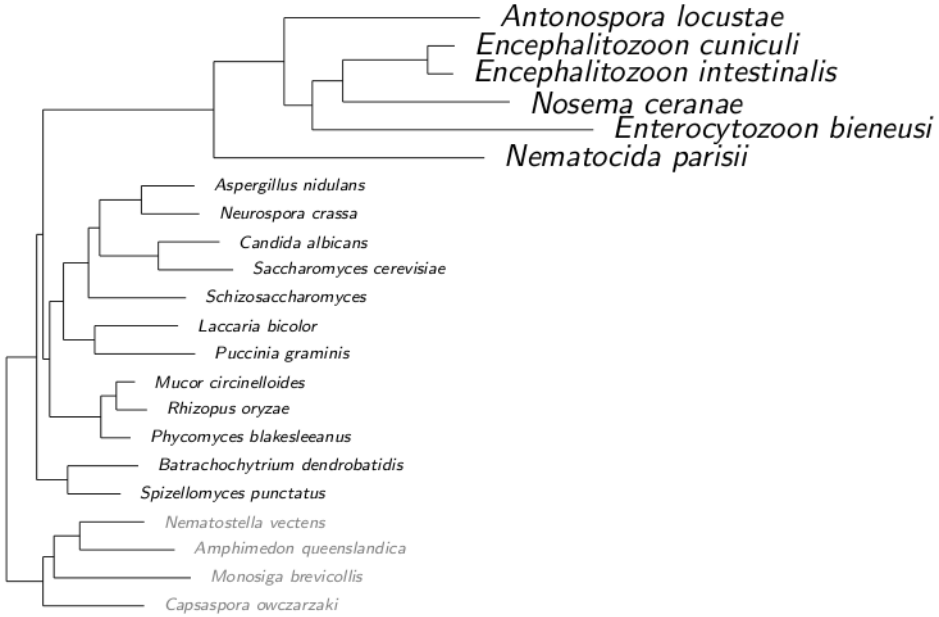


Figure 13 – Arbre phylogénétique enraciné utilisé avec le jeu de données de microsporidies. Le groupe des microsporidies correspond aux noms d'espèces écrits avec de grands caractères, les autres champignons sont écrits avec de plus petits caractères de couleur noir et le *outgroup* est en gris. Les longueurs de branches ont été estimées *a posteriori* sous notre modèle MM6$_\Gamma$

Mesostigma

Mesostigma est une algue verte biflagellée énigmatique au point de vue évolutif. Son positionnement exact relatif au groupe des Streptophyta et à celui des Chlorophyta est encore aujourd'hui un mystère (Lemieux et al. 2007, Rodriguez-Ezpeleta et al. 2007). Nous avons fait quelques tests avec topologie libre sous notre modèle Markov-modulé appliqué sur un jeu de données de séquences protéiques plastidiques concaténées incluant celle de *Mesostigma*. Mais les résultats obtenus ne sont pas très concluants et ne seront donc pas présentés dans le présent travail. L'incorporation ce deuxième jeu de données à notre phase expérimentale n'est en fait nécessaire que pour évaluer l'impact qu'il a sur le comportement de notre modèle comparativement à celui de microsporidies.

Le jeu de données en question est le résultat de la concaténation de 51 protéines chloroplastiques pour un total de 10 137 positions alignées (Rodriguez-Ezpeleta et al. 2007). Rappelons que les protéines codées par les chloroplastes sont majoritairement membranaires et sont donc plutôt hydrophobes. Puisque que les profils des états cachés *HSp6* ont été estimés avec des alignements de protéines nucléaires, ceux-ci ne sont par conséquent pas très ajustés au jeu de données *Mesostigma*.

L'arbre phylogénétique utilisé (figure 14) est constitué de 28 espèces réparties comme suit : 2 groupes de Viridiplantae (Sreptophyta, Chlorophyta), 1 groupe de Rhodophyta et 1 groupe de Glaucophyta. Il est le résultat d'un échantillonnage de topologies de la distribution postérieure par MCMC sous le modèle GTR+Γ_4. Le consensus moyen a été obtenu avec le programme **bpcomp** inclus dans PhyloBayes-MPI version 1.4 à partir de deux chaînes MCMC indépendantes en respectant les recommandations des concepteurs.

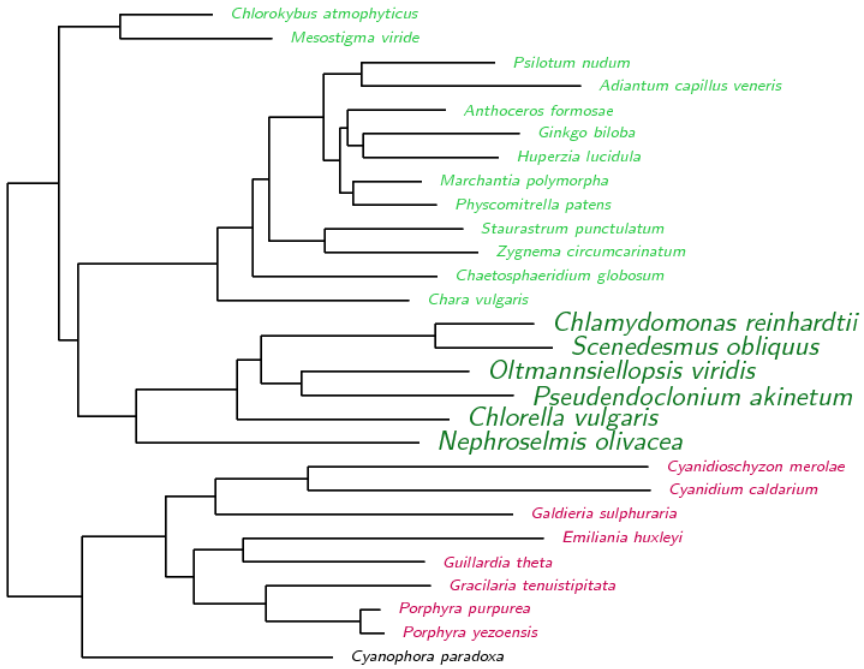


Figure 14 – Arbre phylogénétique postérieur consensus enraciné, inféré avec le modèle GTR+Γ_4, utilisé avec le jeu de données *Mesostigma*. Il est constitué des 4 groupes suivants : Streptophyta, Chlorophyta, Rhodophyta, Glaucophyta.

4 RÉSULTATS

4.1 Estimations *a posteriori*

Nous avons tout d'abord appliqué notre modèle sur un jeu de données obtenu de la littérature (Capella-Gutirrez et al. (2012)) visant à analyser le cas du positionnement phylogénétique des microsporidies. Précisons que tous les résultats qui seront présentés ci-dessous ont été obtenus sous la contrainte d'une topologie fixe. Rappelons également que le modèle Markov-modulé dont on parle ici s'appuie sur un ensemble de 6 matrices empiriques de substitution entre acides aminés (ECG6) estimées sur une base d'alignements de séquences de protéines nucléaires (Groussin et Lartillot, in prep) en tant que modèle de mélange. Dans le cas présent, ces $G = 6$ matrices empiriques représentent les 6 états cachés (*HSp6*) de nos modèles Markov-modulés MM6_Γ et MM6_{cov}. Les modèles homologues MM3_Γ et MM3_{cov} utilisent quant à eux les états cachés *HSp3* correspondant aux états cachés 1, 5 et 6 de *HSp6* (figure 11).

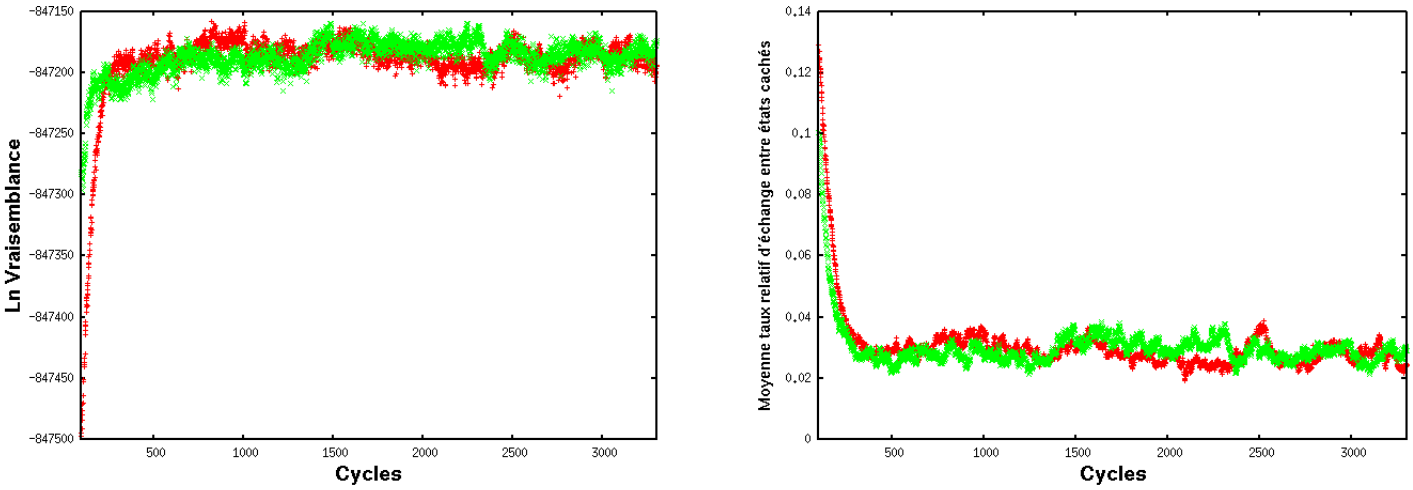
Sont donc estimés directement à partir de l'alignement analysé ici (celui du jeu de données de microsporidies) les taux relatifs d'échange entre états cachés ($\eth_{kl}$) et les fréquences d'équilibre de ces états cachés (η^k) ; ces paramètres constituent les paramètres cachés du modèle. Les paramètres auxiliaires estimés sont les longueurs de branches et le paramètre de forme (α) de la D-Gam. Dans la suite, les différentes estimations ponctuelles qui seront présentées sont en termes de moyennes *a posteriori*.

Notons que les 6 profils des états cachés *HSp6* ne représentent pas les différentes classes biochimiques d'acides aminés aussi bien que les profils des 11 classes stables obtenues par Lartillot & Philippe (2004) avec leur modèle CAT. Néanmoins, la teneur élevée en A, F, G, L et V des états cachés 3 et 4 nous permet de caractériser ceux-ci comme étant plutôt hydrophobes. L'état caché 6 est quant à lui celui dont la variance des fréquences d'équilibre est la plus faible ($\sigma^2 = 0.00063$). Il est donc compatible avec les sites soumis à peu de contraintes sélectives et environnementales. Le profil de l'état caché 2 à une variance plus élevée mais ne semble pas lui non plus avoir de préférence biochimique particulière. Dû à leur teneur élevée respective en proline (P) et en glycine (G), il nous est permis de croire que les états cachés 1 et 5 pourraient être préférentiellement adoptés par des sites interférant avec la formation d'hélices alpha (Pace & Scholtz 1998). Soulignons ici que le nombre de profils n'est pas optimal puisque les différents sites protéiques ne sont pas en réalité limités à un spectre de processus substitutionnels aussi restreint. Le profil 2 est par exemple probablement une conséquence de cette restriction. Il serait plus raisonnable qu'il soit scindé en deux profils biochimiquement opposés ; l'un plutôt hydrophobe riche en F, G, Y et un second polaire/chargé riche en D, E, S, T.

La raison principale pour laquelle nous avons restreint à 6 le nombre d'états cachés est que c'est le compromis que nous avons choisi pour réduire le temps nécessaire à la convergence des chaînes MCMC. Rappelons qu'avec les réglages implantés dans notre échantillonneur Bayésien, le temps nécessaire pour générer un seul point d'une chaîne sous MM6_Γ avec 72 processeurs parallélisés est en moyenne de 28 minutes (comparativement à 4 minutes sous MM3_Γ) sur le jeu de données de microsporidies. Le calcul de la

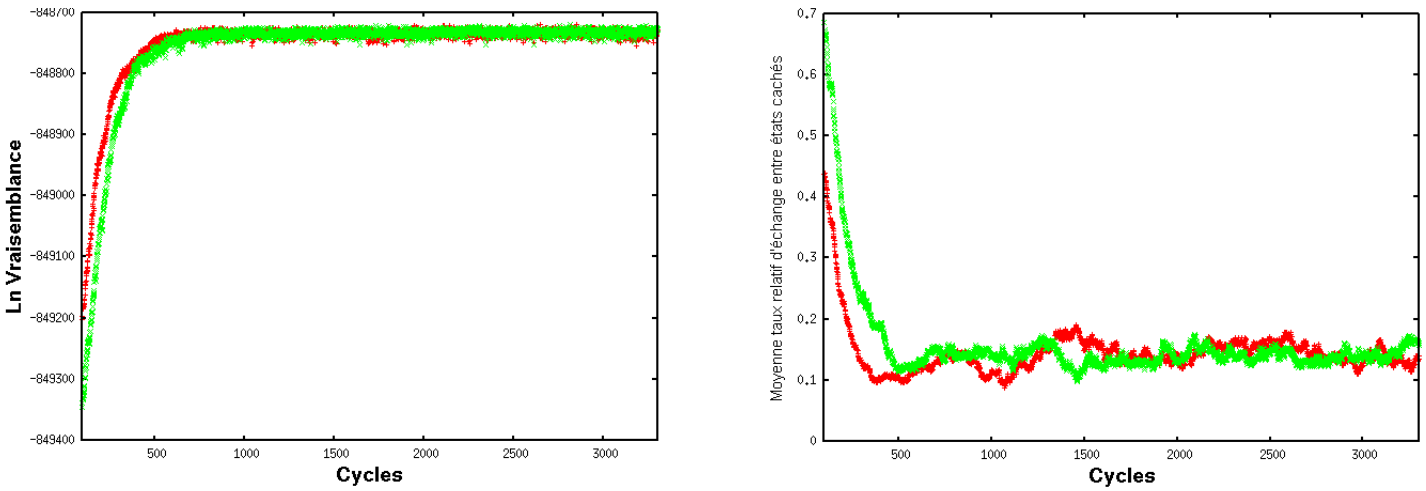
vraisemblance avec une matrice modulée d'ordre 120 est relativement long. Il l'est davantage dans le cas d'un processus non-réversible en temps étant donné qu'il nécessite de calculer l'exponentielle de la matrice modulée pour chaque catégorie de vitesse d'évolution de la D-Gam par une méthode d'approximation numérique (équation 30) beaucoup plus longue que la méthode par diagonalisation.

Les figures 15a et 15b présentent respectivement l'évolution des valeurs de log-vraisemblance et de la moyenne des $\eth_{kl}$ ($\bar{\eth}_{kl}$) durant l'élongation de deux chaînes MCMC indépendantes sous le modèle $MM6_\Gamma$. Les figures 16a et 16b présentent les résultats équivalents obtenus avec $MM3_\Gamma$. Dans les deux cas, nous observons que la convergence de la log-vraisemblance est atteinte après environ 500 cycles. Nos résultats démontrent que cette vitesse de convergence est en fait directement liée à celle du paramètre $\eth$ qui nécessite aussi 500 cycles pour atteindre la phase de convergence. Les autres paramètres tels la longueur de l'arbre, α et l'entropie sur les η^k ont des phases de *burn-in* plus courtes soit respectivement 20, 50 et 150 cycles (résultats non présentés). Précisons ici que globalement la convergence des chaînes MCMC sont très approximatives. Elles sont néanmoins montrées ici dans la mesure où elles nous semblent représenter des résultats qualitativement fiables, même s'ils sont plus discutables sur un plan quantitatif.



(a) **(b)**

Figure 15 – Suivis des log-vraisemblances de θ (a) et de la moyenne des $\eth_{kl}$ ($\bar{\eth}_{kl}$) (b) durant l'élongation de deux chaînes MCMC indépendantes sous le modèle $MM6_\Gamma$ sur le jeu de données de microsporidies.



(a) **(b)**

Figure 16 – Suivis des log-vraisemblances de θ (a) et de la moyenne des $\eth_{kl}$ ($\bar{\eth}_{kl}$) (b) durant l'élongation de deux chaînes MCMC indépendantes sous le modèle $MM3_\Gamma$ sur le jeu de données microsporidies.

Les moyennes estimées *a posteriori* des $\bar{\eth}_{kl}$ (correspondant à la figure 15b) sous MM6_Γ sont de 0.029 pour les deux chaînes MCMC. Leurs intervalles de crédibilité (IC) 95% sont de [0.023, 0.036]. Celles estimées sous MM3_Γ (correspondant à la figure 16b) sont près de cinq fois supérieures ; 0.14 avec des IC 95% respectifs de [0.10, 0.17] pour la chaîne 1 et de [0.12, 0.16] pour la chaîne 2. Notre modèle Markov-modulé semble donc effectivement révéler que les sites protéiques changent de régime de substitution au cours du temps. Ce qui suggère que le modèle CAT serait insuffisant pour expliquer les données empiriques. C'est à dire que dans la mesure ou la densité de probabilité présente un mode très marqué pour des valeurs positives de taux relatifs d'échange moyens entre états cachés, cela semble exclure fortement la configuration ou ces taux moyens sont égaux à 0.

Une raison qui explique pourquoi les $\bar{\eth}_{kl}$ sont plus élevés sous MM3_Γ comparés à ceux estimés sous MM6_Γ est que le taux relatif d'échange entre les états cachés 1 et 5 est beaucoup plus élevé sous MM3_Γ ; $\eth_{15} = 0.39$ versus 0.08 (figures 20c versus 20a). De plus, cette augmentation de $\eth_{15}$ est si prononcée que la somme des $\eth_{kl}$ sous MM3_Γ est supérieure à celle sous *HSp6*. Le modèle semble donc réagir fortement à la perte des taux relatifs $\eth_{12}$, $\eth_{13}$, $\eth_{14}$, $\eth_{52}$, $\eth_{53}$ et $\eth_{54}$ au profit de $\eth_{15}$.

Ces $\bar{\eth}_{kl}$ ne sont en réalité pas très révélateurs du comportement du processus modulant sous-jacent par rapport au processus de substitution entre acides aminés. De plus, compte tenu de leurs moyennes *a priori* (les $\eth_{kl}$ sont tirés d'une distribution exponentielle de moyenne G/S ; équation 41) comparativement à celles estimées *a posteriori* respectivement sous MM6_Γ et sous MM3_Γ, il semble visiblement qu'il y ait un défaut de normalisation.

Nous avons comparé autrement les processus modulants en estimant pour chacun des modèles les taux de transition entre états cachés divisés par les taux de transition totaux

tels que définis par la matrice Q ($P_j(Q)$; équation 60). Il s'agit en fait de l'espérance du nombre de transitions entre états cachés par rapport au nombre total d'événements. Les figure 17 et 18 présentent les distributions postérieures de ces $P_j(Q)$ estimées respectivement sous $MM6_\Gamma$ et $MM3_\Gamma$ (avec deux chaînes MCMC indépendantes dans tous les cas). La moyenne sous $MM6_\Gamma$ est de 0.035 (IC 95% = [0.030, 0.043] et [0.029, 0.039]) comparativement à 0.05 (IC 95% = [0.038, 0.06] et [0.040,0.068]) sous $MM3_\Gamma$. Cet écart est moins important par rapport à ce que nous avons vu pour les $\bar{\eth}_{kl}$. Mais il nous indique néanmoins qu'effectivement le nombre d'états cachés a un impact sur la fréquence des modulations par rapport à celle des transitions entre états observés. La raison demeure toujours inconnue pour l'instant mais des efforts seront éventuellement mis pour tenter de régler ce problème de normalisation.

Donc globalement, que ce soit sous $MM3_\Gamma$ ou sous $MM6_\Gamma$, la proportion de transitions entre états cachés est beaucoup plus faible que celle entre acides aminés. Sous les deux modèles, *a posteriori*, le nombre moyen de substitutions par sites (calculé à partir des histoires substitutionnels tirées de la distribution postérieure) à travers l'arbre du jeu de données de microsporidies est de 20. Ce qui en ressort finalement est que partant de la racine, il y a en moyenne à chaque site 1 modulation pour ces 20 substitutions entre acides aminés. Nous pouvons également conclure en comparant les distributions postérieures et *a priori* (figures 17 et 18) que cette dernière est non-informative. Les valeurs de $P_j(Q)$ postérieures semblent donc effectivement extraites du signal phylogénétique contenu dans les données.

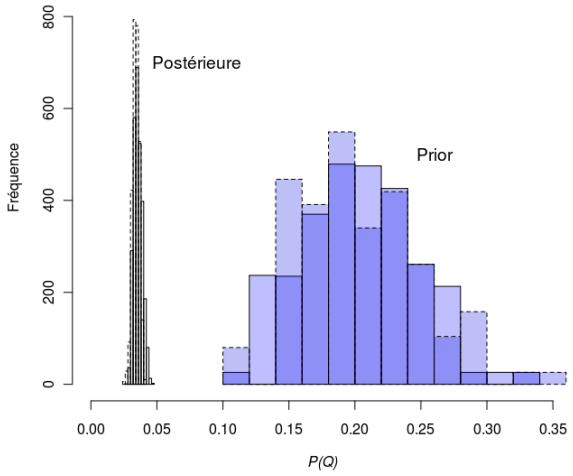


Figure 17 – Distributions postérieures et *a priori* des $P_j(Q)$ estimées à partir de deux chaînes MCMC indépendantes sous $MM6_\Gamma$ sur le jeu de données de microsporidies.

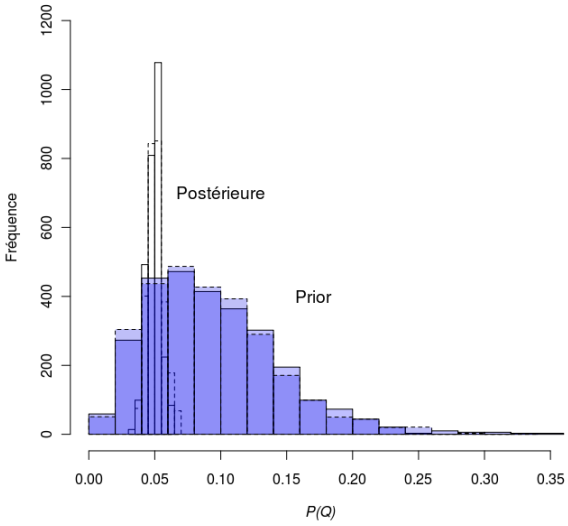


Figure 18 – Distributions postérieures et *a priori* des $P_j(Q)$ estimées à partir de deux chaînes MCMC indépendantes sous $MM3_\Gamma$ sur le jeu de données de microsporidies.

Les estimés des η^k sont aussi très reproductibles tels qu'illustrés à la figure 19. Autant sous $MM6_\Gamma$ que sous $MM3_\Gamma$, l'état caché 6 est nettement dominant suivi de l'état caché 1. Ces fréquences d'équilibre ont un impact important sur les taux de modulation instantanés (équation 14) impliquant les états cachés 1 et 5 sous $MM6_\Gamma$ tels que présentés à la figure 20b. Étant données les faibles fréquences d'équilibre de ces états cachés comparativement à celle de l'état caché 6, les taux instantanés $C^{1,5}$ et $C^{5,1}$ sont beaucoup plus faibles que $C^{2,6}$ et $C^{4,6}$.

(a) **(b)**

Figure 19 – *Logos* représentant les η^k des états cachés *HSp6* (a) et *HSp3* (b) sous les modèles $MM6_\Gamma$ et $MM3_\Gamma$ appliqués sur le jeu de données de microsporidies avec deux chaînes MCMC indépendantes. La hauteur des nombres est proportionnelle à la fréquence d'équilibre de l'état caché correspondant.

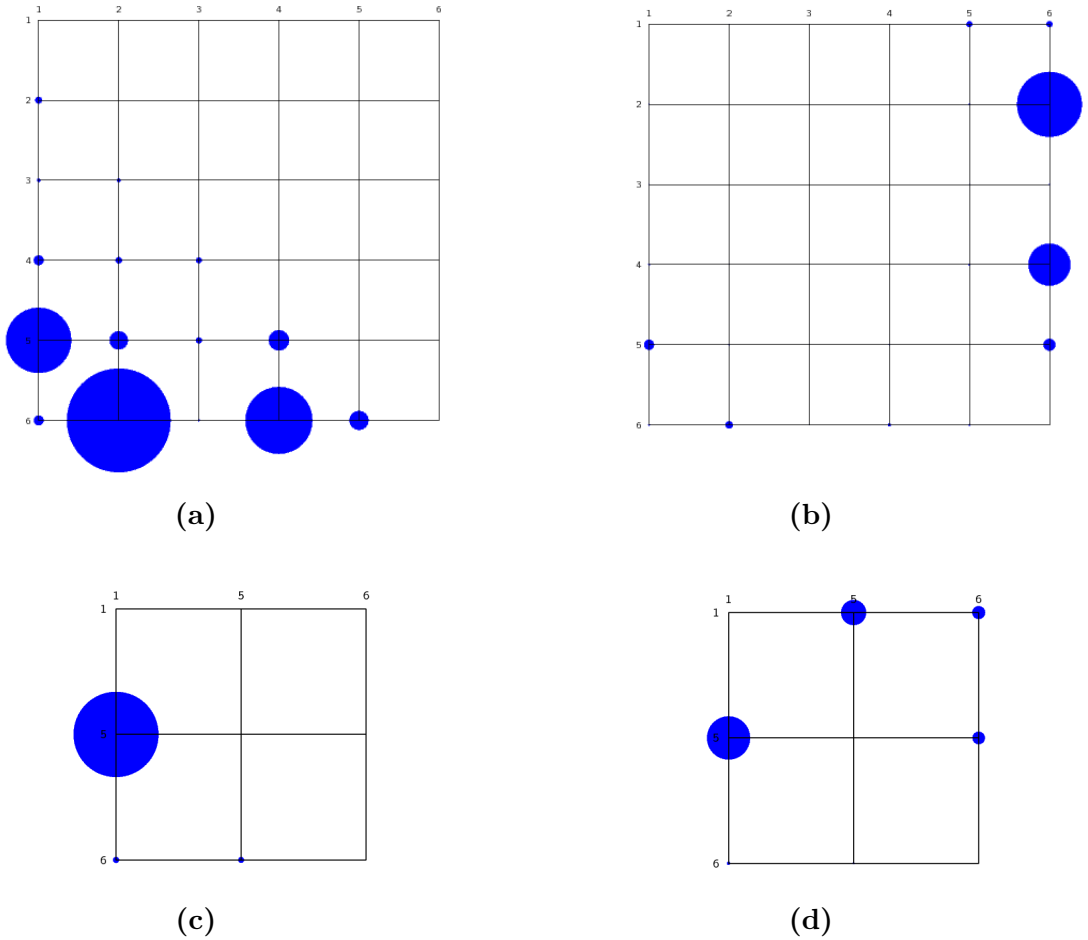
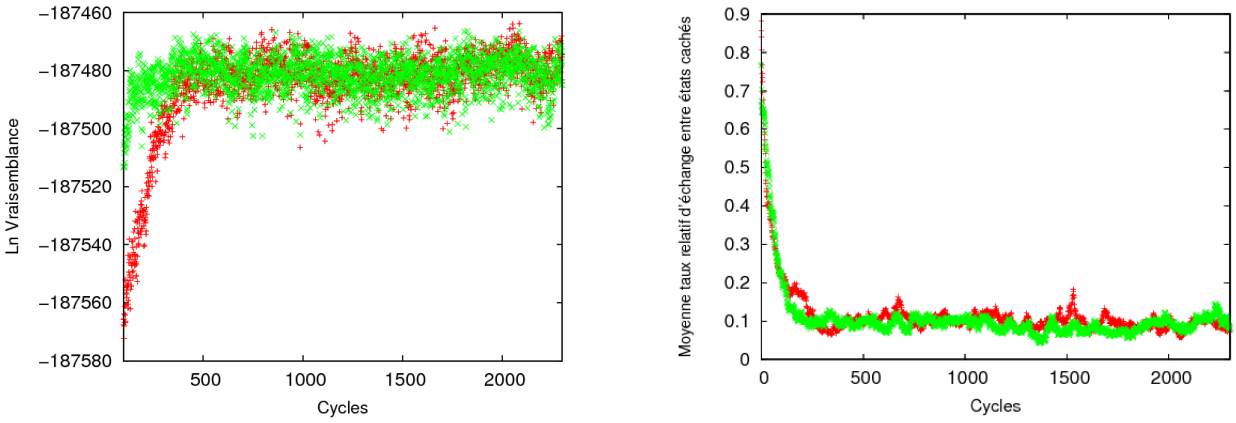


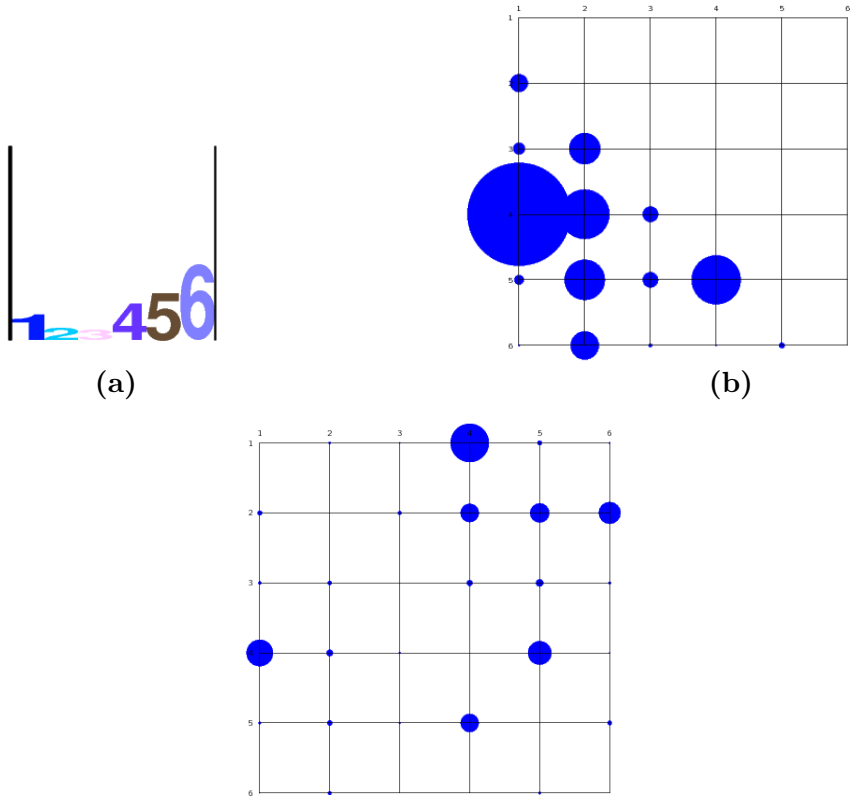
Figure 20 – *Bubble plots* représentant les $\eth_{kl}$ (**a** avec $MM6_\Gamma$ et **c** avec $MM3_\Gamma$) et les $C^{k,l}$ (**b** avec $MM6_\Gamma$ et **d** avec $MM3_\Gamma$) estimés avec les modèles $MM3_\Gamma$ et $MM6_\Gamma$ appliqués sur le jeu de données de microsporidies. Les dimensions des bulles n'ont pas été normalisées entre $MM6_\Gamma$ et $MM3_\Gamma$ puisque l'intention ici n'est que de permettre la comparaison entre les $\eth_{kl}$ et les $C^{k,l}$ d'un même modèle.

Nous avons également évalué le comportement de notre modèle MM6_Γ sur le jeu de données chloroplastiques *Mesostigma*. Nous avons constaté que la convergence des log-vraisemblances après 400 cycles est aussi dépendante de celle des $\bar{\eth}_{kl}$ (figures 21a et 21b). Les $\bar{\eth}_{kl}$ estimés sont plus élevés que ceux estimés avec le jeu de données de microsporidies ; 0.09 avec des IC 95% de [0.064, 0.13] pour la première chaîne et de [0.062, 0.12] pour la seconde. En fait, les $\eth_{kl}$ estimés à partir des deux jeux de données n'ont rien en commun (figure 20a versus 22b). Les taux élevés $\eth_{26}$, $\eth_{46}$ et $\eth_{15}$ estimés à partir du jeu de données de microsporidies sont au contraire presque nuls avec celui de *Mesostigma*. C'est entre les états cachés 1 et 4 que le taux relatif d'échange est le plus élevé avec ce jeu de données. Celui-ci est approximativement 4 fois supérieur au taux le plus élevé estimé à partir du jeu de données de microsporidies ; $\eth_{14} = 0.46$ avec *Mesostigma* versus $\eth_{26} = 0.12$ avec microsporidies.



(a) **(b)**

Figure 21 – Suivis des log-vraisemblances de θ (a) et de la moyenne des $\eth_{kl}$ ($\bar{\eth}_{kl}$) (b) durant l'élongation de deux chaînes MCMC indépendantes sous le modèle $MM6_\Gamma$ sur le jeu de données *Mesostigma*.



(a) **(b)**

(c)

Figure 22 – Paramètres cachés estimés à partir du jeu de données *Mesostigma* sous $MM6_\Gamma$. (a) *Logos* représentant les η^k. (b) *Bubble plots* représentant les $\eth_{kl}$. (c)*Bubble plots* représentant les $C^{k,l}$.

Les η^k des états cachés *HSp6* estimés avec le jeu de données *Mesostigma* sont aussi différents de ceux estimés avec le jeu de données de microsporidies ; notons les fréquences d'équilibre plus élevées des états cachés 4 et 5 à la figure 22a. Mais mise à part ces différences, il est néanmoins intéressant de noter que le taux de modulation instantané le plus élevé estimé avec le jeu de données de *Mesostigma* ($C^{1,4} = 0.072$) est relativement rapproché du plus élevé estimé avec le jeu de données de microsporidies ($C^{2,6} = 0.077$). Ceci s'explique par la fréquence d'équilibre plus élevée de l'état caché 6 avec le jeu de données de microsporidies (0.63 versus 0.37 avec le jeu de données *Mesostigma*) qui compense pour l'important écart entre les taux $\eth_{14}$ (*Mesostigma*) et $\eth_{26}$ (microsporidies) décrit plus haut.

Les paramètres cachés de notre modèle semblent donc sensibles au jeu de données utilisé. Rappelons que le jeu de données *Mesostigma* est chloroplastique et donc riche en protéines transmembranaires hydrophobes. Le jeu de données de microsporidies est quant à lui nucléaire et ainsi davantage ajusté avec les états cachés *HSp6* estimés sur des protéines cytosoliques. Il est fort probable que cet écart de comportement de $MM6_\Gamma$ soit effectivement dû au contenu en acides aminés très différents entre les deux jeux de données.

Contrairement aux matrices de taux relatifs d'échange entre acides aminés estimées empiriquement (par exemples, les matrices JTT, WAG, LG et mtRev) avec lesquelles on peut observer des propensions de remplacement entre des acides aminés ayant des propriétés biochimiques similaires, cela ne semble pas nécessairement transposable aux états cachés. En effet, notons par exemple le taux d'échange faible entre les états cachés 3 et 4 (avec les deux jeux de données, figures 20a et 22b) qui ont pourtant des profils relativement similaires qualitativement. Et inversement, avec le jeu de données *Mesostigma*, le taux d'échange élevé entre l'état caché 1 riche en proline (et plutôt négatif) et l'état caché 4 avec des préférences surtout hydrophobes (figure 22b). La difficulté principale à laquelle nous sommes confrontée avec un tel modèle est que les états cachés ne sont pas des états identifiables de façon univoque comme le sont les 20 acides aminés. Nous ne pouvons qu'estimer le profil de chacun des états cachés ainsi que leur nombre. Ainsi, contraindre un modèle à inférer des taux d'échange entre des états cachés incertains tant au point de vue quantitatif que qualitatif à partir d'un jeu de données spécifique est beaucoup moins exact que le même exercice avec les acides aminés. Nous avons tenté d'estimer des profils d'états cachés par MCMC ainsi que leur nombre par *reversible jump* (Green 1995) mais avec peu de succès étant donné la surdose de paramètres libres causant des problèmes de convergence (résultats non présentés).

Tel que déjà mentionné, il existe un écart important dans les fréquences d'équilibre des états cachés 4 et 5 dépendamment qu'ils aient été estimées avec le jeu de données *Mesostigma* ou avec celui de microsporidies. Ce qui n'est pas le cas pour les états cachés 1, 2, 3 et 6 qui ont des fréquences d'équilibre très reproductibles d'un jeu de données à l'autre (figure 19a versus 22a). Cette observation est directement liée à la similarité entre les taux de modulation instantanés **nets** allant vers et partant respectivement de chacun

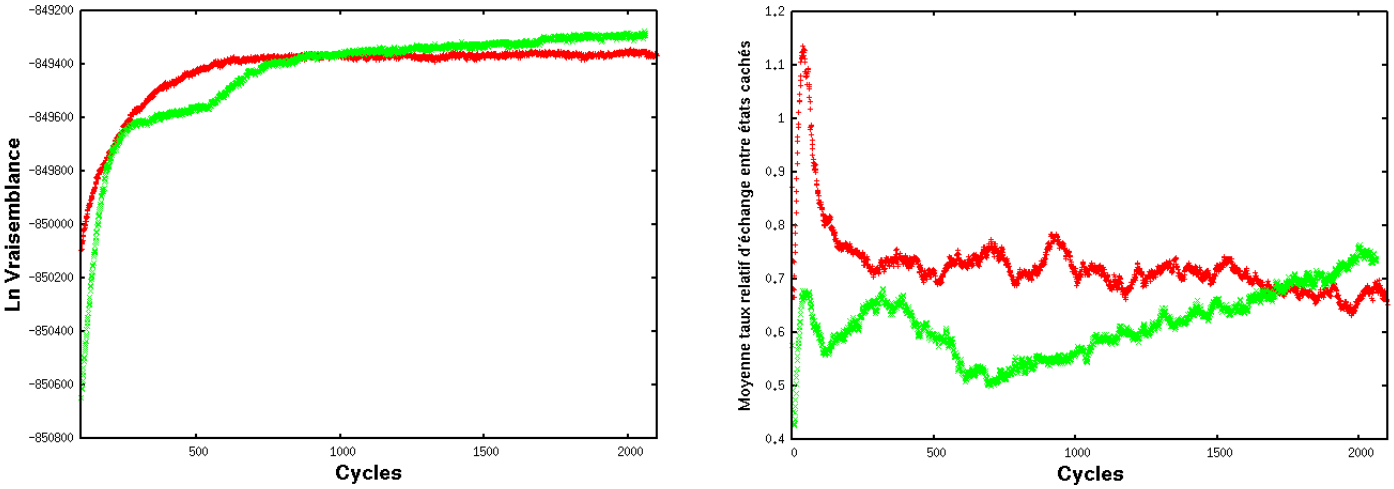
de ces états selon $\sum_{l=1}^{6} C^{k,l}$ - $\sum_{l=1}^{6} C^{l,k}$, $k \in 1,2,3,6$ (figures 20b et 22c). Par exemple, avec le jeu de données de microsporidies, la fréquence d'équilibre élevée de l'état caché 6 est causée par un très fort taux de modulation allant vers ($\Uparrow \sum_{l=1}^{5} C^{l,6}$) et un faible taux partant de cet état. Tandis qu'avec le jeu de données *Mesostigma*, la dominance de l'état caché 6 est plutôt due à un très faible taux partant ($\Downarrow \sum_{l=1}^{5} C^{6,l}$) de cet état comparativement à ceux partant des autres états cachés.

L'état caché 6 est celui rappelons-le dont l'entropie sur les fréquences d'équilibre des acides aminés est la plus faible. C'est aussi l'état caché qui a la fréquence d'équilibre la plus élevée dans tous les cas présentés jusqu'à maintenant. Nous avons suggéré que celui-ci pourrait biologiquement être compatible avec des sites soumis à peu de contrainte sélective. Des sites qui n'ont en fait aucune propension particulière. Mais souvenons-nous également que nos modèle $MM3_{\Gamma}$ et $MM6_{\Gamma}$ ont leur limite. Ils ne possèdent pas suffisamment de composantes pour décrire la richesse réelle des processus substitutionnels sous-jacents. Le profil de l'état caché 6 servirait ainsi de profil 'poubelle' dans un tel contexte restreint . C'est à dire que tous les processus de substitutions qui ne sont pas bien accommodés par les états cachés 1 ou 2 sous $MM3_{\Gamma}$, ou 1 à 5 sous $MM6_{\Gamma}$, modulent par défaut vers l'états cachés 6 qui généralise un ensemble très hétérogène mais aussi très large de modes substitutionnels.

Les estimés *a posteriori* que nous venons de présenter sont ceux obtenus sous l'hypothèse que les processus de substitution Markov-modulés sont combinés avec une variation de vitesse d'évolution à travers les sites modélisée par une distribution gamma. Les configurations $MM3_{\Gamma}$ et $MM6_{\Gamma}$ conçues pour saisir des signaux temporels de changements dans les préférences en acides aminés n'accommodent aucunement le phénomène d'hétérotachie. Nous avons assumé jusqu'à présent que chacun des sites conservait la même vitesse de substitution à travers l'arbre. Il est toutefois aujourd'hui clairement établi que la variation des vitesses d'évolution site-spécifiques est largement répandue dans les jeux de données (Lopez et al. 2002) et qu'il est donc important de ne pas négliger cette réalité évolutive lors de l'élaboration d'un nouveau modèle phylogénétique.

Nos modèles $MM3_{cov}$ et $MM6_{cov}$ ont été construits de manière à ce qu'ils puissent prendre en charge simultanément à la fois les modulations entre les processus de substitution profil-spécifiques et celles entre les vitesses d'évolution. La dimension hétérotache ici est relativement simpliste en ce sens qu'elle n'accommode que deux vitesses de d'évolution ; une vitesse nulle (OFF) et l'autre non-nulle (ON). Nous nous sommes en fait basé sur le premier modèle mathématique proposé pour les processus covarion, soit celui de Tuffley & Steel (1998). Donc sous $MM3_{cov}$ et $MM6_{cov}$, la modélisation de RAS n'est pas prise en charge de manière explicite avec une distribution gamma. Elle est remplacée par un état caché supplémentaire invariant (que nous avons noté **0**) à l'intérieur duquel le taux d'évolution est nul.

Nous pouvons remarquer sur les figures 23 (sous $MM6_{cov}$) et 24 (sous $MM3_{cov}$) que les $\eth\bar{}_{kl}$ estimés sous nos modèle de type covarion sont visiblement non convergents et qu'ils sont plus élevés que ceux estimés avec MM_{Γ}. Sous $MM6_{cov}$; 0.70 (IC 95% [0.65, 0.73]) pour la première chaîne et 0.65 (IC 95% [0.56, 0.73]) pour la seconde. Sous $MM3_{cov}$, ils sont encore plus élevés comme c'est le cas avec la configuration MM_{Γ} ; 4.3 (IC 95% [3.9, 4.5]) pour la première chaîne et 4.5 (IC 95% [3.9, 4.5]) pour la seconde. Les proportions espérées de transitions entre états cachés (calculées à partir des entrées de la matrice Q ($P(Q)$)) sont présentées sous forme de distribution postérieure à la figure 25. Leurs valeurs suggèrent qu'il y existe une proportion d'événements de modulation très élevée sous ce cette configuration. Sous $MM6_{cov}$; 0.49 (IC 95% [0.47, 0.50]) avec la première chaîne et 0.48 (IC 95%

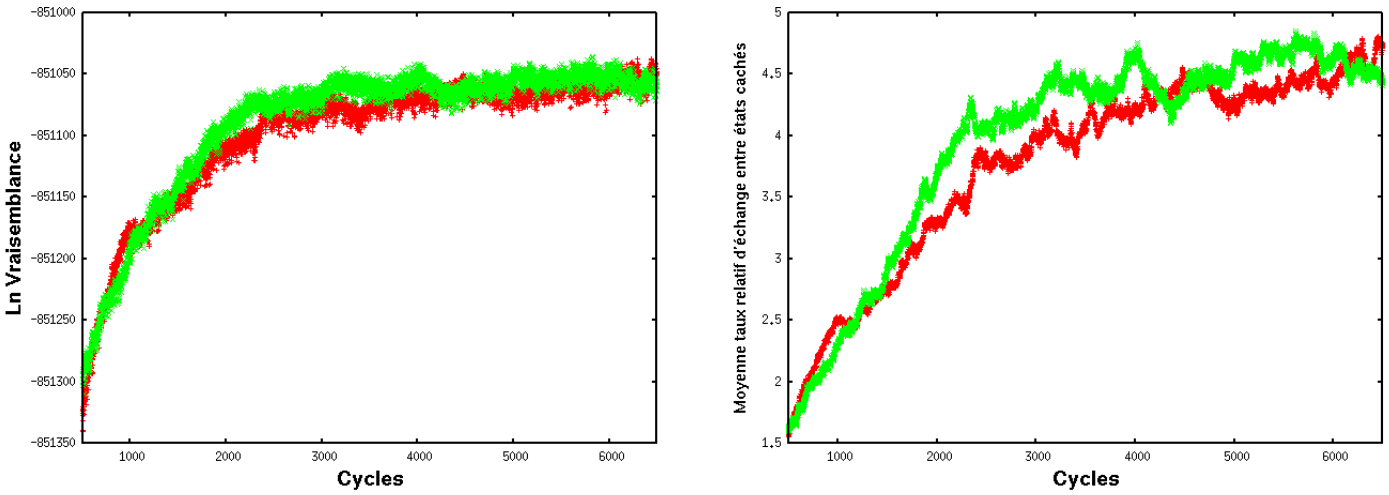


(a) **(b)**

Figure 23 – Suivid des log-vraisemblances de θ (a) et de la moyenne des $\eth_{kl}$ ($\eth\bar{}_{kl}$) (b) durant l'élongation de deux chaînes MCMC indépendantes sous le modèle $MM6_{cov}$ sur le jeu de données de microsporidies.

[0.45, 0.52]) avec la deuxième chaîne. Et sous $MM3_{cov}$; 0.79 (IC 95% [0.78, 0.80]) avec la première chaîne et 0.80 (IC 95% [0.79, 0.81]) avec la seconde. Le nombre moyen *a posteriori* de transitions entre acides aminés par site à travers l'arbre étant également de 20, il y a donc parallèlement autant d'événements modulants sous $MM6_{cov}$ comparativement à 80 sous $MM3_{cov}$. L'impact du nombre d'états cachés sur la fréquence des modulations par rapport à celle des transitions entre état observés (dû au défaut de normalisation) est donc beaucoup plus important avec la configuration covarion.

L'insuffisance de la convergence des chaînes MCMC est encore plus évidente sous MM_{cov}. Néanmoins, qualitativement, il est possible d'avoir une idée générale du comportement du modèle comparativement au modèle MM_{Γ} en ce sens que l'incorporation d'un état caché OFF amplifie substantiellement le phénomène de modulation.



(a) **(b)**

Figure 24 – Suivi des log-vraisemblances de θ (a) et de la moyenne des $\eth_{kl}$ ($\bar{\eth}_{kl}$) (b) durant l'élongation de deux chaînes MCMC indépendantes sous le modèle $MM3_{cov}$ sur le jeu de données de microsporidies.

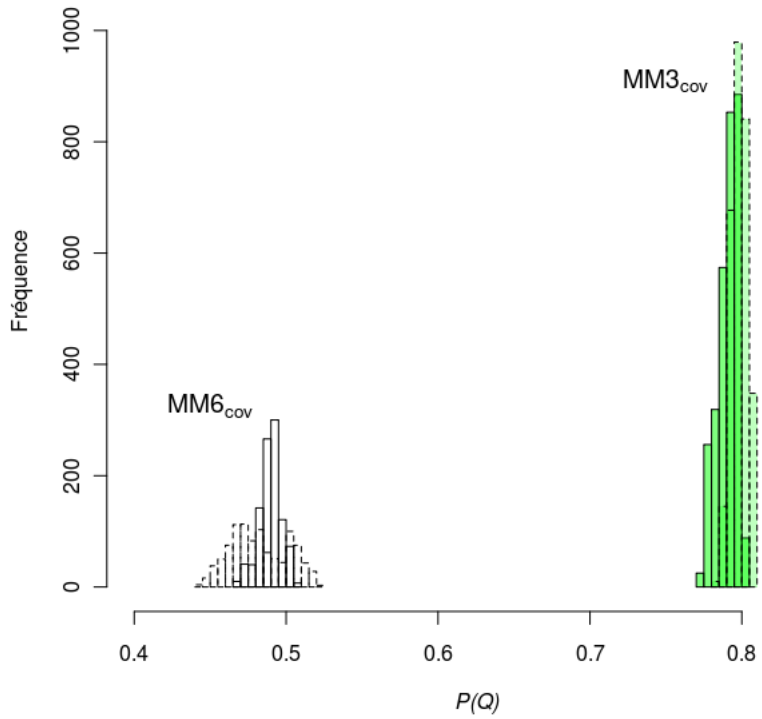
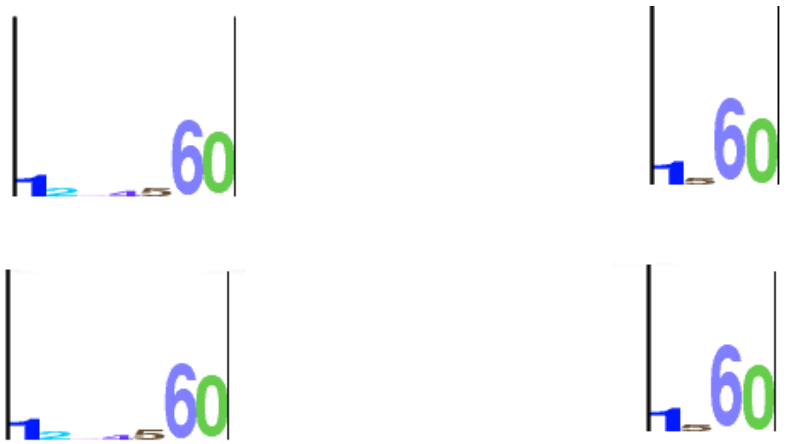


Figure 25 – Distributions postérieures des $P_j(Q)$ estimées à partir de deux chaînes MCMC indépendates sous $MM6_{cov}$ et $MM3_{cov}$ sur le jeu de données de microsporidies.

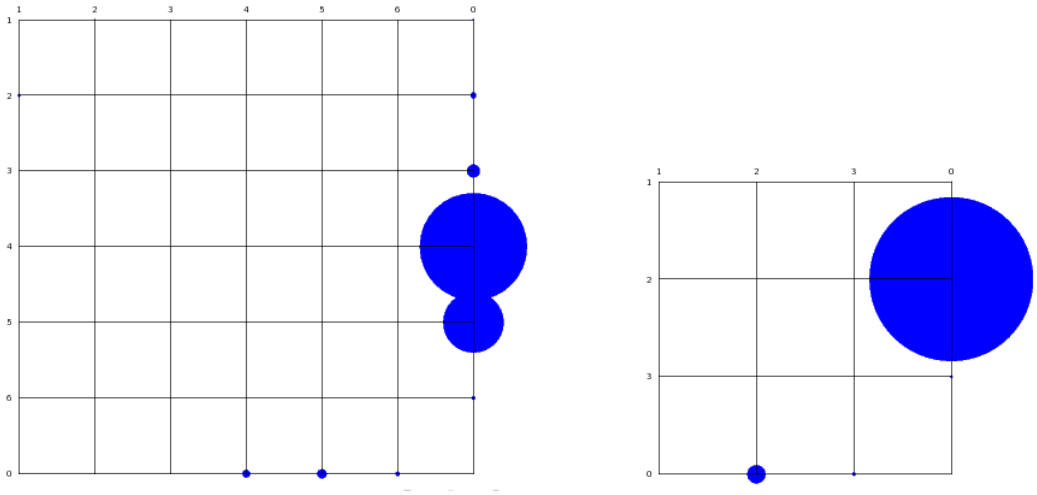
Excluant la fréquence d'équilibre de l'état caché **0**, les η^k estimés sous les modèles MM_{cov} sont identiques à ceux estimés sous les modèles MM_{Γ} (figure 26 versus figure 19). Notons la fréquence d'équilibre élevée de l'état caché **0** autant sous $MM6_{cov}$ que sous $MM3_{cov}$. Cela s'explique par l'imposant taux total de transition allant vers l'état caché **0** (figure 27) comparativement aux autres taux ; par exemple sous $MM6_{cov}$, $\Uparrow \sum_{l=1}^{6} C^{l,0}$ avec une forte contribution provenant des taux $C^{4,0}$ et $C^{5,0}$. Cette fréquence d'équilibre n'est néanmoins pas aussi élevée que celle de l'état caché 6 puisque le taux total partant de l'état caché **0** est aussi relativement élevé (tout en étant cependant de beaucoup inférieur au taux total allant vers l'état caché **0**) comparativement aux taux entre les autres cachés . En résumé, ces résultats démontrent donc que sous MM_{cov} il y a prévalence des modulations de type hétérotache comparativement à celles de type profil-spécifique.



(a) **(b)**

Figure 26 – *Logos* représentant les η^k des états cachés *HSp6* (a) et *HSp3* (b) sous les modèles $MM6_{cov}$ et $MM3_{cov}$ appliqués sur le jeu de données de microsporidies avec deux chaînes MCMC indépendantes. La hauteur des nombres est proportionnelle à la fréquence d'équilibre de l'état caché correspondant.

Les sérieux problèmes de convergence observés sous $MM6_{cov}$ seraient dus, selon les *bubble plots* présentés aux figures 28a et 28b, aux taux relatifs $\eth_{40}$ et $\eth_{50}$. Ceux-ci sont non congruents entre les deux chaînes MCMC ; 7.1 versus 3.1 pour $\eth_{40}$ et 4.6 versus 9.1 pour $\eth_{50}$. L'absence de l'état caché 4 sous $MM3_{cov}$ semble résoudre cette non congruence du fait que les $\eth_{50}$ estimés sont très rapprochés (figures 29a et 29b) ; 25.6 avec la première chaîne et 25.5 avec la seconde.



(a) **(b)**

Figure 27 – *Bubble plots* représentant les taux de transition $C^{k,l}$ estimés sous les modèles $MM6_{cov}$ (a) et $MM3_{cov}$ (b) appliqués au jeu de données de microsporidies. Les dimensions des bulles n'ont pas été normalisées entre $MM6_{cov}$ et $MM3_{cov}$ puisque l'intention ici n'est que de permettre la comparaison entre les $C^{k,l}$ d'un même modèle.

Afin de vérifier si la configuration MM_{cov} modifiait les taux relatifs d'échange entre les 6 premiers états cachés par rapport à ceux observés sous MM_{Γ}, nous avons présenté sur les figures 28c et 28d une vue rapprochée sur ces taux en filtrant ceux avec l'état caché **0**. Nous pouvons constater une meilleure congruence entre les deux chaînes mais surtout la disparition de la dominance du taux $\eth_{26}$ observée sous $MM6_{\Gamma}$ au profit de $\eth_{12}$. Les taux $\eth_{15}$ et $\eth_{46}$ sont aussi significativement réduits. Le même exercice effectué avec $MM3_{cov}$ révèle que, tout comme sous $MM6_{\Gamma}$, c'est entre les états cachés 1 et 5 que le taux relatif d'échange est le plus élevé (figures 29c et 29d). $\eth_{15}$ semble cependant être celui qui empêche la congruence des $\bar{\eth}_{kl}$ sous ce modèle ; 0.256 pour la première chaîne et 0.112 pour la seconde.

Nos résultats suggèrent donc que l'état caché **0** semble interférer avec la capture des signaux de modulation entres les différents processus de substitution profil-spécifiques. Le taux total de transition élevé vers l'état caché **0** démontre cependant que l'introduction d'un paramètre lié au phénomène d'hétérotachie semble requis pour expliquer les données empiriques. Mais la configuration de ce modèle telle que nous l'avons conçue montre des aspects qui sont quelque peu surprenants et difficilement interprétables sur un plan biologique. En particulier, le modèle semble exagérer les transitions ON/OFF.

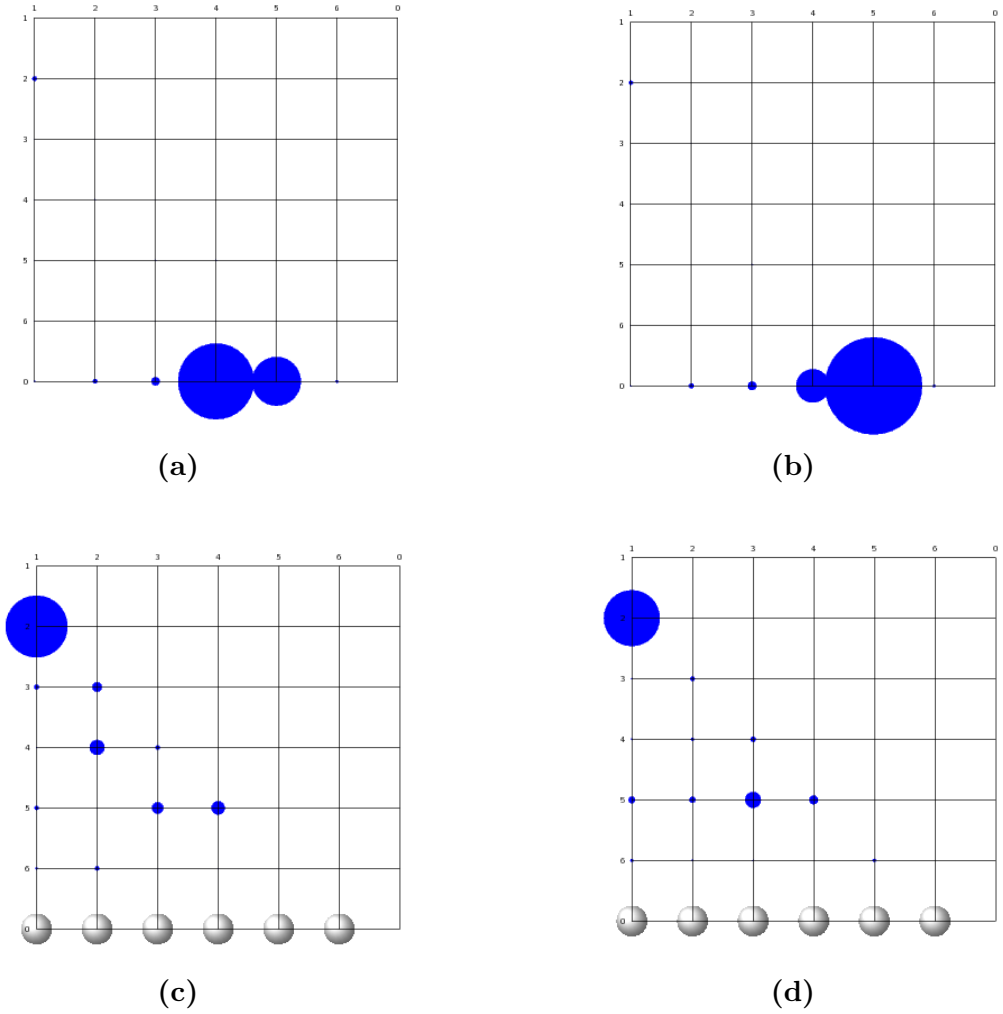


Figure 28 – *Bubble plots* représentant les $\eth_{kl}$ estimés sous le modèle $MM6_{cov}$ appliqué au jeu de données de microsporidies. **a** et **c** présentent ceux estimés avec la première chaîne MCMC et **b**, **d** ceux avec la seconde. **c** et **d** sont issus de la filtration des taux d'échange avec l'état caché **0** sur les figures **a** et **b** respectivement.

Un autre point faible du modèle MM_{cov} est que la matrice modulante Q^{cov} perd la mémoire de l'état caché modulant vers l'état caché **0**. Par exemple, prenons un site hydrophobe sur une branche donnée de l'arbre sous un certain taux d'évolution qui devient soumis à une forte pression sélective au temps t_1. Selon notre modèle, lorsqu'au temps t_2 cette pression est assouplie, les probabilités de revenir au processus de substitution hydrophobe ou de moduler vers un autre régime disponible sont exactement les mêmes que si le site était initialement sous un régime à caractère hydrophile. Ce qui, biologiquement, est également difficilement interprétable. Les modulations avec l'état caché **0** interféreraient avec les signaux de divergences functionnelles potentiels. Ce qui pourrait expliquer les différences entre les taux relatifs d'échange estimés sous le modèle $MM6_{\Gamma}$ et les taux équivalents (après filtration des taux $\eth_{k0}$) sous le modèle $MM6_{cov}$.

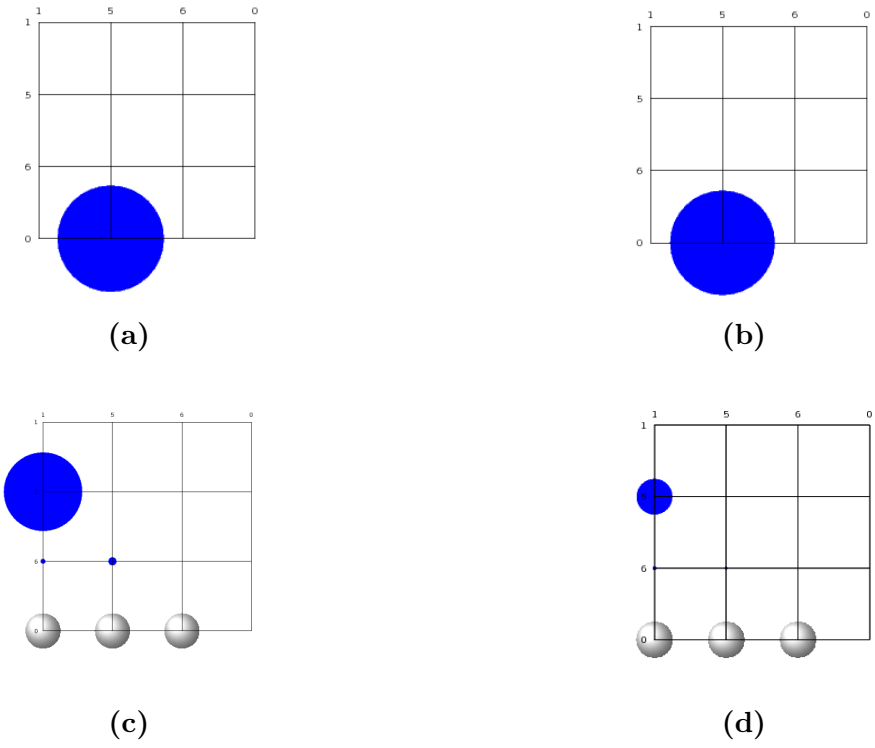


Figure 29 – *Bubble plots* représentant les $\eth_{kl}$ estimés sous le modèle $MM3_{cov}$ appliqué au jeu de données de microsporidies. **a** et **c** présentent ceux estimés avec la première chaîne MCMC et **b**, **d** ceux avec la seconde. **c** et **d** sont issus de la filtration des taux d'échange avec l'état caché **0** sur les figures **a** et **b** respectivement.

4.2 Comparaison de modèles par validation croisée

La comparaison des modèles quant à leur ajustement aux données est importante. En ML, on utiliserait à cette fin des méthodes telles LRT (*Likelihood Ratio Test*)(Neyman & Pearson 1933), AIC (*Akaike Information Criterion*)(Akaike 1977) et BIC (*Bayesian Information Criterion*)(Schwartz 1978). Ici en Bayésien, classiquement on voudrait comparer les modèles (M_i) sur la base de leur vraisemblance marginale définie par

$$P(D \mid M_i) = \int_{\theta} P(D \mid \theta, M_i) P(\theta \mid M_i) d\theta. \tag{63}$$

Ce qui en fait implique de calculer des facteurs de Bayes (Jeffreys 1935 ; Jeffreys 1961 ; Jaynes 2003) entre modèles. Toutefois, les facteurs de Bayes sont potentiellement sensibles à la prior (Sinharay & Stern 2002) et notoirement difficiles à calculer.

Une alternative : la validation croisée (Smyth 2000). Elle consiste principalement à entraîner un modèle sur un jeu de données d'apprentissage D_l puis ensuite à évaluer la vraisemblance d'un jeu test D_t sous le même modèle entraîné. En Bayésien, la puissance prédictive de M est définie comme étant la probabilité de D_t sachant D_l ou encore

l'espérance de la vraisemblance conditionnelle à D_t par rapport à la distribution postérieure sous D_l et M définie par

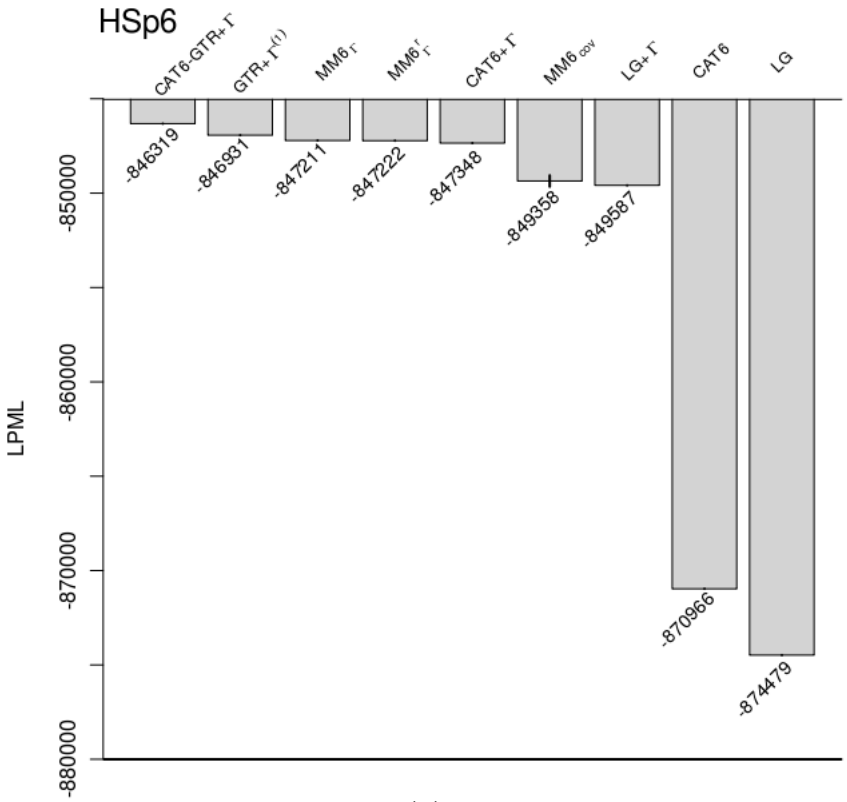
$$p(D_t \mid D_l, M) = \int_\theta p(D_t \mid \theta, M) p(\theta \mid D_l, M) d\theta. \tag{64}$$

Cette façon de faire à l'avantage de corriger automatiquement pour les aspects de surapprentissage.

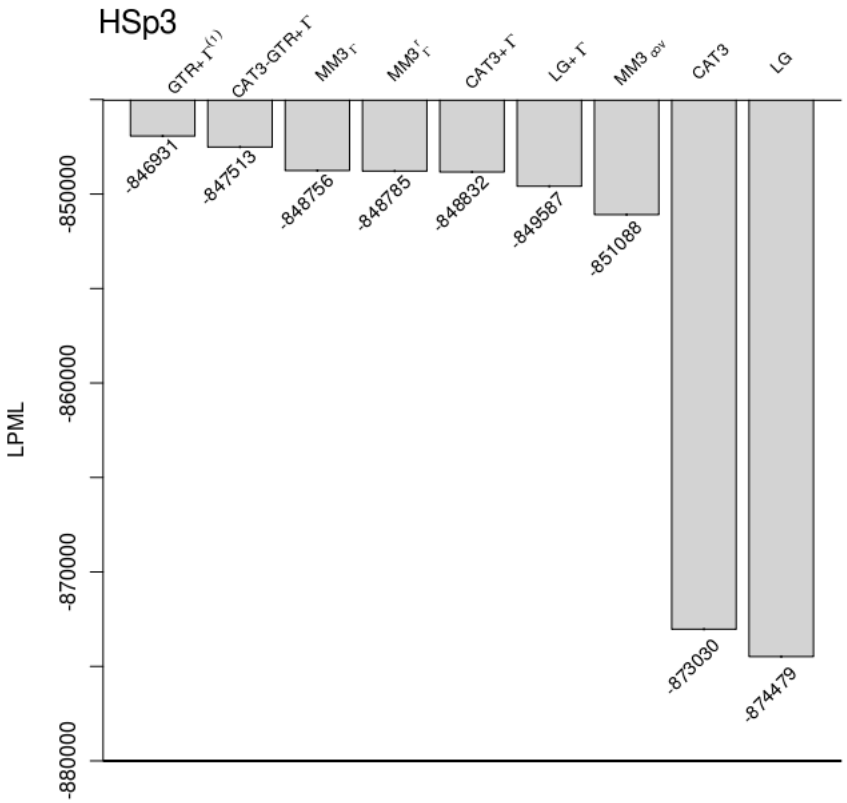
Deux approches sont possibles pour scinder le jeu de données en D_t et D_l ; soit couper en deux moitiés de manière aléatoire ou utiliser la méthode du *leave-one-out* (LOOCV). En phylogénie moléculaire, cette dernière consiste à mesurer la valeur d'ajustement à une colonne C_i (D_t) de l'alignement d'un modèle entraîné à partir de toutes les autres colonnes. L'algorithme associé au LOOCV doit boucler ainsi à travers les colonnes de façon à ce que toutes aient été traitées comme étant D_t.

Récemment, Lewis et al. (2013) proposent une méthode LOOCV combinée à une approche d'échantillonnage préférentielle. La stratégie consiste à premièrement récupérer par MCMC un échantillon de taille A ($\theta^1, \theta^2, \ldots, \theta^A$) de la distribution postérieure conditionnelle à l'ensemble des données. Le score de validation croisée au site C_i (CPO_i) est obtenu en calculant pour chacun des points de l'échantillon la contribution de ce C_i à la vraisemblance totale. Plus précisement, l'on estime l'inverse de la moyenne harmonique postérieure de la vraisemblance au site C_i telle que définie par l'équation 46. C'est l'approche que nous avons adoptée ici et réimplémentée dans notre programme.

La valeur LPML telle que définie par l'équation 48 résulte de la sommation des log CPO_i. Rappelons qu'elle est la mesure utilisée pour évaluer la performance d'un modèle. Les LPML obtenus pour chacun des modèles Markov-modulés appliqués sur le jeu de données de microsporidies sont présentées sous forme d'histogrammes aux figures 30a (avec 6 états cachés) et 30b (avec 3 états cachés).



(a)



(b)

Figure 30 – Évaluation de la performance des modèles Markov-modulés avec la méthode CPO. LPML obtenus sous les modèles $MM3_\Gamma/MM6_\Gamma$ et $MM3_{cov}/MM6_{cov}$ et sous d'autres modèles de référence exécutés sur le jeu de données de microsporidies. **a** avec les états cachés *HSp6* et **b** avec les états cachés *HSp3*. (1)Rappelons que les fréquences d'équilibre des acides aminés sont libres sous GTR+Γ.

Soulignons tout d'abord que la liberté des fréquences d'équilibre des acides aminés et de leurs taux relatifs d'échange sous le modèles GTR+Γ contribue substantiellement à l'augmentation de la CPO. D'autre part, il est intéressant de noter que nos modèles Markov-modulés non-réversibles ($MM3_\Gamma$, $MM6_\Gamma$) et réversibles ($MM3^r_\Gamma$, $MM6^r_\Gamma$) sont légèrement plus performants que les modèles CAT3+Γ et CAT6+Γ. D'après les calculs faits à partir de l'équation 49, les gains moyens en *CPO* par site sont de 1.0054× avec $MM6_\Gamma$ et de 1.0049× avec $MM6^r_\Gamma$. Considérant que les modèles CAT3+Γ et CAT6+Γ sont en fait les équivalents des modèles $MM3_\Gamma$, $MM6_\Gamma$, $MM3^r_\Gamma$ et $MM6^r_\Gamma$ dont les $\eth_{kl}$ sont tous égaux à 0, cela suggère que les données empiriques semblent effectivement avoir besoin d'une dimension modulée des processus de substitution profil-spécifiques pour être expliquées. De plus, les résultats suggèrent un meilleur ajustement lorsque l'on contraint le processus modulé à ce qu'il soit non-réversible en temps ; $\text{LPML}_{MM6_\Gamma} > \text{LPML}_{MM6^r_\Gamma}$ et $\text{LPML}_{MM3_\Gamma} > \text{LPML}_{MM3^r_\Gamma}$.

La figure 31 montre le gain progressif en ajustement avec le degré d'hétérogénéité spatiale (entre les sites) et temporelle (à travers le temps). Notons le gain considérable lorsque le nombre d'états cachés du modèle MM_Γ passe de 3 à 6 (+1 545). Mais surtout, l'incorporation de RAS est de loin ce qui permet le plus important gain en ajustement (+24 892). L'observation de ces deux contributions hétérogènes est d'ailleurs en accord avec les résultats obtenus par Whelan (2008).

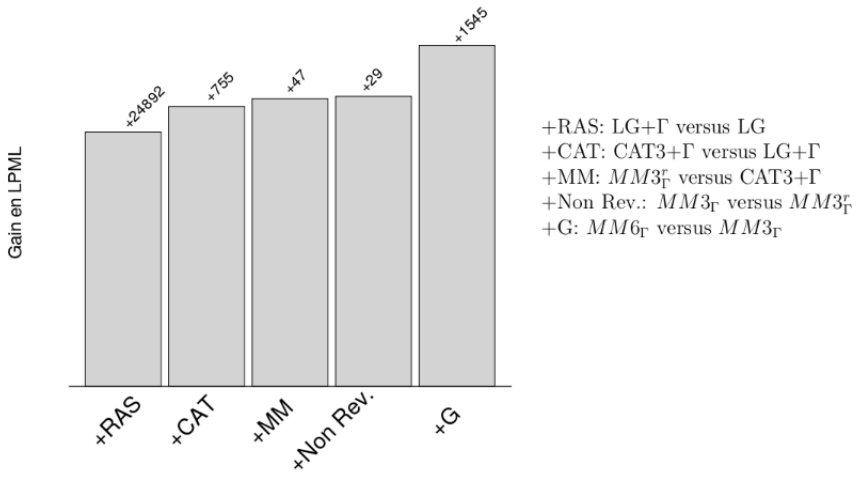


Figure 31 – Gain progressif en ajustement aux données avec l'augmentation du degré d'hétérogénéité du modèle.

Notons la perte importante de performance en absence de RAS. La valeur LPML chute de -23 618 sous CAT6 (CAT6 versus CAT6+Γ), de -24 198 sous CAT3 (CAT3 versus CAT3+Γ) et de -24 892 sous le modèle LG (LG versus LG+Γ). Le modèle LG est d'ailleurs de loin le modèle le moins performant avec une valeur LPML de -874 479. La figure 32 montre que la configuration $MM6_{cov}$ n'est pas suffisamment hétérogène en temps pour reproduire l'effet RAS.

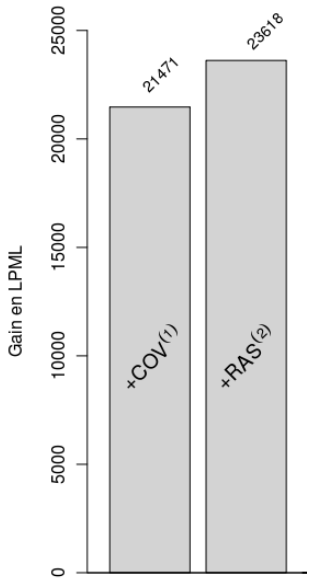


Figure 32 – Accommodement partiel de l'hétérogénéité des taux d'évolution entre sites par le modèle $MM6_{cov}$. $^{(1)}$($\text{LPML}_{MM6_{cov}}$ - $\text{LPML}_{\text{CAT6}}$) - ($\text{LPML}_{MM6_{\Gamma}}$ - $\text{LPML}_{\text{CAT6+}\Gamma}$). $^{(2)}$$\text{LPML}_{\text{CAT6+}\Gamma}$ - $\text{LPML}_{\text{CAT6}}$.

Afin d'augmenter la significativité statistique de ces résultats, nous en convenons qu'il serait souhaitable de refaire le même exercice sur d'autres jeux de données. Par exemple, sur des jeux de données de protéines mitochondriales (caractérisées par un degré élevé d'hydrophobicité). Il serait également intéressant de faire des essais sur de plus petits jeux de données ou en réduisant le nombre d'états observés à 4 avec des profils nucléotidiques. Et un mélange plus riche en états cachés permettrait-il une performance plus persuasive de nos modèles Markov-modulés par rapport aux modèles CAT3+Γ et CAT6+Γ ?

4.3 Histoires substitutionnelles et tests prédictifs *a posteriori*

Nous avons jusqu'à maintenant estimé *a posteriori* les différents paramètres cachés définissant nos modèles Markov-modulés ; ð, η, $P(Q)$. Nous avons également mesuré leur performance et pu noter que la modélisation du processus modulé semble effectivement améliorer l'ajustement aux données empiriques. Dans ce qui suit, nous allons regarder d'un peu plus près le comportement des modèles $MM6_{\Gamma}$ et $MM6_{cov}$. Cela fait partie de la démarche exploratoire naturelle à suivre pour mieux évaluer si le modèle se comporte oui ou non comme nous l'avions espéré au départ. Nous allons plus précisement regarder ce qui se passe exactement le long des branches, à savoir la séquence des événements de substitutions et de modulations à travers l'arbre.

Le concept de divergence fonctionnelles (DF) fait référence principalement au processus évolutif qui mène à une modification ou à la perte de fonction d'un gène. Les DF peuvent survenir suite à une duplication de gène (Henikoff & al. 1997 ; Li 1983). Dans ce cas, l'un des deux gènes paralogues conserve la fonction ancestrale alors que l'autre est libre d'évoluer à sa guise. Les DF peuvent également cibler certains gènes suite à des changements environnementaux majeurs. Les gène orthologues adaptent alors leur fonction à la nouvelle dynamique biochimique nécessaire à la survie de l'organisme. Le phénomène de l'endosymbiose est d'ailleurs à l'origine de DF chez certaines bactéries (Caffrey et al. 2012 ; Toft et al. 2009) et chez les microsporidies (Keeling & Fast 2002).

Les DF impliquent les positions qui ont un rôle important à jouer dans la fonction de la protéine ; liaison de cofacteurs et de substrats divers (protéines, acides nucléiques, composés organiques, ...) ou nécessaires au maintien de la structure. D'où l'intérêt d'identifier ces dites positions. Dans la littérature, on parle de deux types de DF (Gu 1999). Le type I résulte d'un changement au niveau de la force de la contrainte évolutive agissant sur un site. Donc, ralentissement ou accélération du taux d'évolution. Le type II concerne quant à lui les changements ancestraux dans les préférences en acides aminés.

Divers classificateurs existent, conçus pour identifier des sites DF de type I et II à partir de différents alignements protéiques. Mais, tel que soulevé par Gaston et al. (2011), la difficulté avec ces classificateurs réside dans l'identification des sites ne contribuant pas à la divergence fonctionnelle (faux positifs et vrais négatifs). Un modèle Markov-modulé tel le notre peut potentiellement être plus spécifique et sensible, et venir complémenter les résultats obtenus avec les classificateurs.

Les événements de modulations détectables à partir d'un jeu de données constitués de protéines de la même famille peuvent effectivement originés d'une duplication ou d'une spécialisation génique. Mais ils peuvent aussi être causés par une altération de l'environnement local de la protéine ou par une modification de son intéraction physique avec une autre protéine comme c'est le cas pour la cytochrome *c* oxydase de la lignée des anthopoïdes (Adkins et al. 1996). Dans la suite, afin de simplifier, le terme DF fera référence à tout événement impliquant des changements dans les préférences en acides aminés. Et

ce, peu importe s'ils sont une conséquence ou non d'une modification dans la fonction de la protéine. Les événements de duplication de gènes ne seront pas mis en cause ici puisque le jeu de données de microsporidies est constitué exclusivement de protéines orthologues.

Le jeu de données de microsporidies est idéal pour évaluer notre modèle quant à sa capacité à y saisir des signaux de DF. La raison est qu'il est maintenant bien établi que les microsporidies sont des champignons ayant évolué rapidement et ayant été la cible de plusieurs changements de régimes substitutionnelles au cours du temps (Keeling & Fast 2002). Notre objectif ici, contrairement aux classificateurs DF, n'est pas d'identifier précisement les positions de l'alignement responsables des DF. Il vise plutôt à vérifier si notre modèle peut mettre en évidence des branches de la lignée des microsporidies qui sont enrichies en DF. Techniquement, dans notre perspective Markov-modulée, ceci se traduit par un nombre proportionnellement plus grand de transitions entre états cachés à travers l'ensemble des sites sur ces dites branches. Cette approche est relativement simple du fait que ces proportions sont calculées à partir des histoires substitutionnelles tirées de la distribution postérieure. Alternativement, il serait possible de raffiner notre modèle en lui ajoutant par exemple des paramètres branche-spécifiques accommodant les proportions de transitions entre états cachés par rapport à celles entre états observés.

4.3.1 Analyses d'histoires substitutionnelles

Pour estimer *a posteriori* sur chacune des branches j la proportion (P_j) de transitions entre états cachés par rapport au nombre total d'événements, nous avons utilisé l'équation 50. Rappelons que cette approche exige premièrement de tirer des histoires substitutionnelles (Ξ) de la distribution postérieure en utilisant la méthode proposée par Nielsen (2002) et adaptée par Rodrigue et al. (2008).

Les résultats que nous avons obtenus sous $MM6_{\Gamma}$ et $MM6_{cov}$ sont présentés en format *Gnuplot* sur la figure 33. Sous le modèle $MM6_{\Gamma}$ (figure 33a), les P_j des deux chaînes convergent (sauf pour trois branches) vers des estimés égals à environ 0.0223. Donc, *a posteriori* cela signifie qu'il y a en moyenne, sur chacune des branches, 1 modulation à toutes les 50 substitutions. Aucune valeur extrême ne semble ressortir ni chez le groupe des microsporidies ($j \geq 32$) ni chez les autres champignons. Nous avons fait le même test avec $MM3_{\Gamma}$ et obtenu un P_j moyen près de celui estimé avec $MM6_{\Gamma}$; soit de 0.0230. Il n'est pas évident d'évaluer la compatibilité de ces P_j avec la réalité évolutive. Mais, il semble néanmoins y exister une certaine cohérence biologique du fait que ces résultats rejettent fortement la probabilité qu'il y ait autant d'événements de modulation que de transitions entre acides aminés.

Les P_j moyens obtenus sous le modèle $MM6_{cov}$ (figure 33b) sont non congruents ; 0.379 versus 0.393. Cela corrèle avec les sérieux problèmes de convergence causés par la dimension hétérotache de ce modèle comme que nous avons pu le constater plus haut. Dû au taux total de transition très élevé allant vers l'état caché **0**, ces P_j moyens sont près de 17 fois supérieurs à ceux estimés sous $MM6_{\Gamma}$. Autrement dit, sous $MM6_{cov}$ il y a

environ 6 modulations entre états cachés pour 10 transitions entre acides aminés. La figure 33c montre la nette amélioration de la congruence suite à la filtration des modulations covarion (voir la figure 12 dans les matériels et méthodes). Ce qui vient confirmer que ce sont effectivement les $\eth_{k0}$ qui causent la non-congruence des P_j. Cette filtration a de plus pour effet d'augmenter la variance des P_j et permet de faire ressortir 5 branches localisées dans le groupe des microsporidies. Les P_j correspondants sont mis en évidence avec des flèches noires et sont associés selon l'arbre présenté à la figure 13 aux branches suivantes ; la longue branche menant au groupe des microsporidies ($j = 32$, $P_j = 0.0815$ et 0.0848) et aux branches externes menant respectivement à *Nematocida* ($j = 33$, $P_j = 0.0957$ et 0.0992), *Enterocytozoon* ($j = 36$, $P_j = 0.0994$ et 0.102), *Nosema* ($j = 38$, $P_j = 0.0764$ et 0.0801) et *Antonospora* ($j = 42$, $P_j = 0.0833$ et 0.0870). Ces résultats supportent donc la nécessité d'introduire une dimension hétérotache dans notre modèle Markov-modulé pour qu'il puisse faire ressortir un signal intéressant provenant du groupe des microsporidies. Et puisque ce sont uniquement les modulations de type covarion qui sont filtrés, il semble que se soit effectivement des signaux de changements dans les processus de substitution profil-spécifiques qui sont saisis par le modèle $MM6_{cov}$.

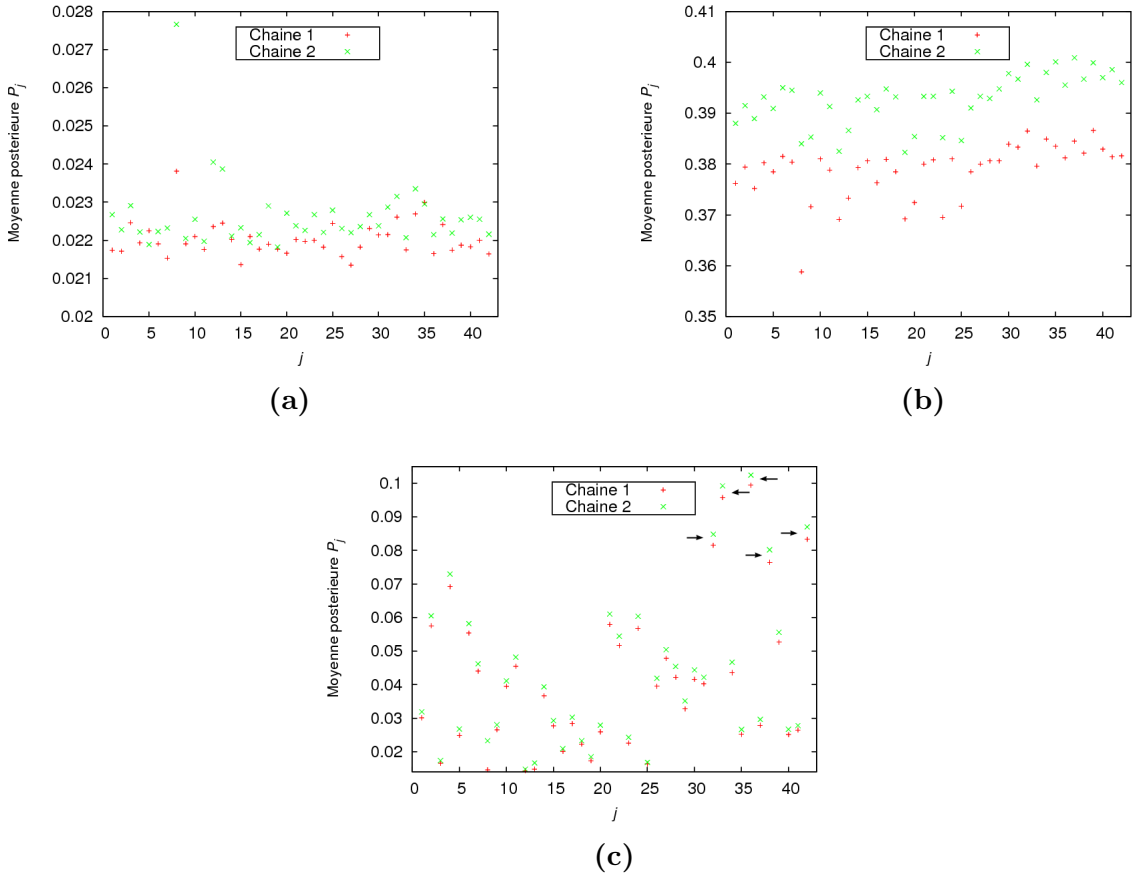


Figure 33 – P_j estimés avec deux chaînes MCMC indépendantes sous $MM6_\Gamma$ (a) et $MM6_{cov}$ (b) pour chacune des branches j de l'arbre du jeu de données de microsporidies. Les P_j de la figure c résultent de la filtration des modulations covarion sous $MM6_{cov}$.

Le tableau 8 présente la corrélation attendue entre les $P_j(Q)$ espérés calculés à partir de la matrice Q et les P_j moyennés sur l'ensemble des branches calculés à partir des histoires substitutionnelles (que nous désignerons par $P_j(\Xi)$ dans la suite). Nous pouvons remarquer que globalement les $P_j(Q)$ sont légèrement supérieurs aux $P_j(\Xi)$ mais qu'ils sont néanmoins relativement concordants. L'impact du nombre d'états cachés sur $P_j(Q)$ (i.e. l'augmentation de $P_j(Q)$ avec la diminution du nombre d'états cachés) est reproduit sur les $P_j(\Xi)$. Rappelons qu'il s'agirait ici d'un phénomène causé par un problème de normalisation des taux de transition entre états cachés. Un problème qui est nécessaire de corriger dans notre prochaine version du modèle.

Tableau 8 – Corrélation entre $P_j(Q)$ et $P(\Xi)$

Modèles	Chaînes	$P_j(Q)$	$P_j(\Xi)$ moyens
$MM6_{\Gamma}$	1	0.035	0.0220
$MM6_{\Gamma}$	2	0.035	0.0226
$MM3_{\Gamma}$	1	0.050	0.0229
$MM3_{\Gamma}$	2	0.050	0.0230
$MM6_{cov}$	1	0.49	0.379
$MM6_{cov}$	2	0.48	0.393
$MM6_{cov}{}^{a}$	1	n/a	0.0409
$MM6_{cov}{}^{a}$	2	n/a	0.0433
$MM3_{cov}$	1	0.79	0.593
$MM3_{cov}$	2	0.80	0.590

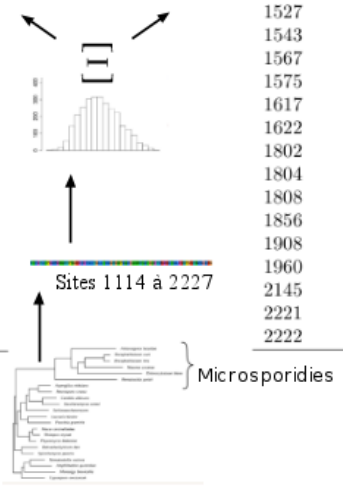
[a] Après filtration des modulations covarion

Nous avons regarder de plus près les histoires substitutionnelles sur la longue branches ($j = 32$) menant au groupe des microsporidies afin d'établir un parallèle général avec les estimés postérieurs des P_j et des $\eth_{kl}$. Pour ce faire, nous avons premièrement tiré de la distribution postérieure, sous $MM6_{\Gamma}$, deux histoires substitutionnelles sur cette branches. Ensuite, nous avons analysé la succession exacte de transitions entre états cachés et entre acides aminés pour environ 1000 positions alignées consécutives.

Les histoires substitutionnelles que nous avons tirées de la distribution postérieure sont présentées sur la figure 34, mais uniquement pour les sites modulants. Tel qu'attendu, les deux histoires sont totalement différentes ; c'est une réalité du contexte Bayésien qui offre une mesure naturelle d'incertitude. Les différents scénarios sont en fait visités à une fréquence qui est fonction de leur probabilité postérieure respective (i.e. $\Xi \sim p(\Xi \mid D, \theta)$). Les statistiques calculées à partir des deux histoires substitutionnelles (tableau 9) sont cependant globalement les mêmes ; densité de modulations par site modulant ($\mathfrak{D} = 1.04$ versus 1.00), densité de sites modulants ($\mathcal{D} = 0.024$ versus 0.024), densité de sites avec au moins une substitution ($d = 0.58$ versus 0.57). Dans les deux cas, nous pouvons également noter que ces résultats sont en parfait accord avec les taux de transition élevés partant

des états cachés 2 et 4 vers l'état caché 6 ($C^{2,6}$ et $C^{4,6}$ sur la figure 20b) ; les transitions $2 \succ 6$ et $4 \succ 6$ totalisent 15 parmi 28.

Sites	Ξ / Echantillon postérieur x
1119	$2 \succ 6[F \to T]$
1184	$6 \succ 1$
1297	$2[Y \to F] \succ 6[F \to M]$
1327	$4[L \to F] \succ 6$
1328	$4 \succ 6[L \to V]$
1331	$2 \succ 6$
1344	$4 \succ 6$
1438	$1 \succ 6$
1440	$6[T \to S] \succ 4[S \to A]$
1457	$6 \succ 4$
1500	$6 \succ 2$
1506	$4[M \to L] \succ 6$
1542	$2 \succ 6$
1623	$2[Q \to T \to A] \succ 6[A \to G]$
1669	$6 \succ 2$
1700	$6[V \to L] \succ 6$
1704	$1 \succ 6$
1773	$2[F \to R] \succ 6$
1791	$5 \succ 6[I \to L \to I]$
1804	$2 \succ 6$
1978	$6[S \to A] \succ 5$
1979	$6[E \to T \to M] \succ 1$
2062	$4 \succ 6[F \to Y \to F \to L]$
2081	$6 \succ 2 \succ 6$
2126	$4[A \to S] \succ 6[S \to T]$
2176	$2 \succ 6$
2182	$6[E \to K] \succ 5[K \to E]$



Sites	Ξ / Echantillon postérieur (x + 20)
1115	$2[E \to D] \succ 6$
1122	$2 \succ 6[D \to E]$
1128	$2 \succ 6$
1192	$2 \succ 6$
1206	$2 \succ 6$
1218	$6 \succ 2$
1242	$6 \succ 2$
1328	$4[L \to F \to S] \succ 6$
1353	$4 \succ 6$
1437	$1 \succ 5[T \to V]$
1464	$5[E \to G] \succ 6$
1523	$6 \succ 1$
1527	$6 \succ 2$
1543	$3 \succ 4$
1567	$1[Y \to F] \succ 5[F \to Y \to H]$
1575	$1 \succ 5$
1617	$1 \succ 6[Q \to L \to Q \to E]$
1622	$2 \succ 6[E \to D]$
1802	$4 \succ 6[L \to M \to I \to V]$
1804	$2 \succ 1[T \to S \to R \to K \to R]$
1808	$2[Q \to S \to D \to N \to D] \succ 6[D \to N]$
1856	$2 \succ 6$
1908	$2 \succ 6$
1960	$2 \succ 6$
2145	$6 \succ 1$
2221	$2[D \to A] \succ 6[A \to S]$
2222	$2[G \to A \to V] \succ 6$

Figure 34 – Deux (x et x+20) histoires substitutionnelles (Ξ) sur la branche à la base du groupe des microsporidies tirées de la distribution postérieure sous le modèle $MM6_{\Gamma}$.

Tableau 9 – Statistiques calculées à partir de deux histoires substitutionnelles tirées de la distribution postérieure sous le modèle $MM6_{\Gamma}$[a]

Ξ/Échantillon postérieur	$\mathfrak{D}^b$	$\mathcal{D}^c$	d^d	$\mathfrak{d}^e$
x	1.04	0.024	0.58	0.45
x + 20	1.00	0.024	0.57	0.54

[a] Calculées à partir des histoires substitutionnelles sur la figure 34.

[b] Densité moyenne de modulations par site modulant.

[c] Densité de sites modulants à travers les 1114 (2227-1114+1) sites successifs.

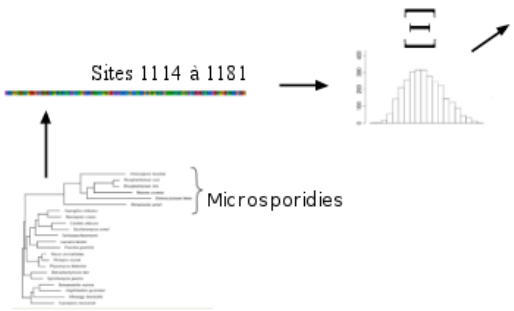
[d] Densité de sites avec substitution(s) à travers les 1114 (2227-1114+1) sites successifs.

[e] Nombre moyen de substitutions par état caché.

Nous avons fait le même exercice avec le modèle $MM6_{cov}$, mais cette fois-ci en comparant une histoire substitutionnelle complète à une seconde filtrée de ses modulations

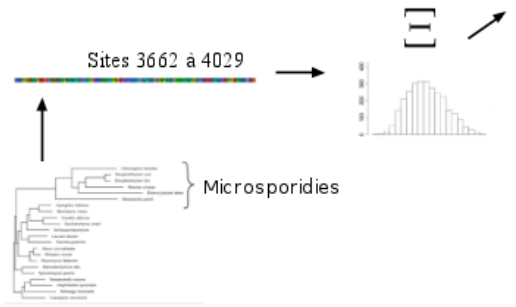
covarion. Les résultats sont présentés respectivement sur les figures 35 et 36. Dans le premier cas, nous pouvons noter qu'effectivement MM_{cov} semble exagérer le phénomène d'hétérotachie. La proportion de modulations avec l'état caché **0** est extrêmement élevée ; parmi les 39 transitions entre état cachés compilées, 37 sont avec l'état caché **0**. Les statistiques calculées (tableau 10) montrent que la densité de sites modulants est près de 16 fois supérieure à celle estimée sous $MM6_{\Gamma}$; $\mathcal{D} = 0.37$ versus 0.024. Ce qui signifie qu'il y a en moyenne 40 sites avec au moins une modulation par 100 sites. Il est également intéressant de noter que ce taux élevé de modulation vers l'état caché **0** a un impact majeur sur le nombre moyen des transitions entre acides aminés par état caché. En moyenne, celui-ci chute environ de moitié sous $MM6_{cov}$ ($\mathfrak{d} = 0.27$ sous $MM6_{cov}$ versus 0.45 pour l'échantillon x et 0.54 pour l'échantillon x+20 sous $MM6_{\Gamma}$). Cette différence pourrait d'ailleurs expliquer partiellement les $P_j(\Xi)$ filtrés plus élevés sur les branches du groupe des microsporidies sous $MM6_{cov}$ (figure 33c). De futures investigations se pencheront sur la question. Néanmoins, rappelons que le nombre moyen de transitions entre acides aminés par site à travers l'arbre est le même sous les deux configurations, soit 20.

Après filtration des modulations covarion, les statistiques sur les histoires de transitions entre états cachés sont beaucoup plus près de celles calculées sous $MM6_{\Gamma}$. Mais notons cependant que parmi les 25 sites modulants filtrés successifs présentés sur la figure 36, 23 transitions entre états cachés profil-spéciques passent par l'état caché **0**. Ce qui encore ici démontre que la dimension hétérotache de MM_{cov} n'est pas tout à fait au point.



Sites	Ξ / Echantillon postérieur x
1114	$6 \succ 0 \succ 4 \succ 0$
1120	$2 \succ 1$
1122	$2 \succ 0 \succ 6$
1124	$2[S \to T] \succ 1[T \succ N]$
1125	$5 \succ 0$
1126	$1[D \to E] \succ 0$
1128	$2[E \to Q \to H \to E] \succ 0$
1131	$6 \succ 0$
1138	$6 \succ 0$
1140	$5 \succ 0 \succ 4 \succ 0 \succ 1$
1141	$5[S \to D] \succ 0 \succ 5$
1142	$0 \succ 1 \succ 0$
1143	$5 \succ 0$
1145	$0 \succ 5$
1149	$0 \succ 1$
1152	$0 \succ 2$
1154	$6[Q \to T] \succ 0 \succ 4$
1155	$0 \succ 4[A \to S] \succ 0$
1160	$2 \succ 0$
1163	$2 \succ 0$
1168	$5 \succ 0$
1172	$5 \succ 0$
1175	$0 \succ 5 \succ 0 \succ 6 \succ 0$
1180	$0 \succ 5 \succ 0$
1181	$0 \succ 5$

Figure 35 – Une histoire substitutionnelle (Ξ) sur la branche à la base du groupe des microsporidies tirée de la distribution postérieure sous le modèle $MM6_{cov}$.



Sites	Ξ / Echantillon postérieur y
3664	$4 \succ 0 \succ 1[S \to P \to A]$
3702	$6 \succ 0 \succ 5$
3710	$6[Q \to N] \succ 0 \succ 4$
3719	$6 \succ 0 \succ 5$
3725	$5 \succ 0 \succ 4$
3730	$5 \succ 0 \succ 4$
3732	$5 \succ 0 \succ 6[Q \to K \to Q \to E]$
3770	$1 \succ 2[S \to T]$
3791	$4 \succ 0 \succ 6$
3796	$6 \succ 0 \succ 4$
3803	$4 \succ 0 \succ 3$
3825	$6[S \to L] \succ 0 \succ 5$
3839	$6 \succ 0 \succ 5$
3858	$6 \succ 2[D \to E]$
3883	$4 \succ 0 \succ 1[V \to T]$
3886	$6 \succ 0 \succ 4$
3898	$5[Q \to K] \succ 0 \succ 4$
3904	$5 \succ 0 \succ 6$
3924	$5 \succ 0 \succ 6[R \to K]$
3952	$5 \succ 0 \succ 4$
3954	$4 \succ 0 \succ 5$
4008	$4 \succ 0 \succ 5$
4025	$5[S \to G] \succ 0 \succ 6$
4028	$4 \succ 0 \succ 6$
4029	$4 \succ 0 \succ 5$

Figure 36 – Une histoire substitutionnelle (Ξ) sur la branche à la base du groupe des microsporidies tirée de la distribution postérieure sous le modèle $MM6_{cov}$ après filtration des modulations covarion.

Tableau 10 – Statistiques calculées à partir d'histoires substitutionnelles tirées de la distribution postérieure sous le modèle $MM6_{cov}$[a]

Ξ	$\mathfrak{D}^b$	$\mathcal{D}^c$	$\mathfrak{d}^g$
Sans filtration des modulations covarion	1.6	0.37[e]	0.27
Avec filtration des modulations covarion	1.0[d]	0.076[f]	0.26

[a] Calculées à partir des histoires substitutionnelles apparaissant sur les figures 35 et 36.
[b] Densité moyenne de modulations par site modulant.
[c] Densité de sites modulants.
[d] Deux modulations successives passant par l'état caché **0** comptent pour une seule modulation (figure 12).
[e] À travers 68 (1181-1114+1) sites successifs.
[f] À travers 368 (4029-3662+1) sites successifs.
[g] Nombre moyen de substitutions par état caché.

4.3.2 Tests prédictifs *a posteriori*

Afin d'évaluer la capacité de nos modèles Markov-modulés à mettre en évidence des branches enrichies en DF, nous avons adopté une seconde approche ; celle du test prédictif *a posteriori* (Meng 1994 ; Gelman et al. 1996). En phylogénie moléculaire, les analyses prédictives *a posteriori* sont souvent utilisées pour vérifier si un modèle est en accord le

réel processus évolutif. Si c'est le cas, les statistiques observées (calculées à partir des données empiriques (D^o)) devraient être indistinguables de celles (statistiques prédictives *a posteriori*) calculées à partir des données simulées (D^s) par le modèle. La diversité nucléotidique moyenne par site (ϵ) est par exemple une statistique qui convient très bien à ce genre d'analyse. Dans un contexte Bayésien, l'idée serait de premièrement calculer ϵ^s pour chacun des D^s simulés avec des θ^a tirés de la distribution postérieure ($D^s \sim p(D^s \mid D^o, M) = p(D^s \mid \theta)$). Et deuxièmement de comparer la distribution prédictive *a posteriori* des ϵ^s avec ϵ^o. Une valeur p prédictive *a posteriori* (valeur p dans la suite) est calculée par exemple en compilant le nombre de fois que $\epsilon^s < \epsilon^o$. Si celle-ci était inférieure à un certain seuil fixé (par exemple 0.05) nous pourrions en conclure qu'il n'y a pas suffisamment d'évidence pour croire que le modèle est compatible avec la réalité évolutive.

Les statistiques utilisées pour notre premier test prédictif *a posteriori* sont les $P_j(\Xi)$ et les z-scores $Z_j(\Xi)$ (équation 55) calculés à partir des histoires substitutionnelles tirées de la ditribution postérieure observée ($\Xi^o \sim p(\Xi^o \mid \theta, D^o)$) et ceux calculés à partir des histoires substitutionnelles tirées de la distribution prédictive *a posteriori* ($\Xi^s \sim p(\Xi^s \mid \theta)$). Pour chaque θ^a ($\theta^a \sim p(\theta^a \mid D^o, M)$), deux histoires substitutionnelles sont simulées ; l'une contrainte aux données empiriques (Ξ^o) et l'autre non contrainte (Ξ^s). Ensuite, pour chacune des branches j, une valeur p est calculée pour les statistiques $P_j(\Xi)$ et $Z_j(\Xi)$. Elle consiste à compter le nombre le nombre de fois (pour l'ensemble des θ^a) que $P_j(\Xi)^s \geq P_j(\Xi)^o$ (ou que $Z_j(\Xi)^s \geq Z_j(\Xi)^o$). Une valeur p inférieure à un certain seuil (voir ci-dessous) suggère un signal significativement plus élevé de DF sur la branche j.

Dans ce qui suit sont présentés premièrement les résultats de congruence des valeurs p sur les $P_j(\Xi)$ (figure 37) et sur les $Z_j(\Xi)$ (figure 38) avec deux chaînes MCMC indépendantes sous les modèles $MM6_\Gamma$ et $MM6_{cov}$ (avec et sans filtration des modulations covarion). Nous avons calculé les moyennes et les variances de ces valeurs p dans chacun des cas (tableau 11).

Le test prédictif *a posteriori* effectué sous $MM6_\Gamma$ est définitivement non concluant. Autant avec la statistique $P_j(\Xi)$ (figure 37a) qu'avec la statistique $Z_j(\Xi)$ (figure 38a), trop de branches ont des valeurs p non congruentes. Nous n'avons pas identifié la cause de ce problème. Mais il est fort probable qu'il soit en lien avec la longueur (3 500 points) et la convergence insuffisante des chaînes MCMC. En pratique, la longueur des chaînes MCMC doit être beaucoup plus longue ($>$ 100 000 points) (Norris 1998). Cela permet entre autres de récupérer ensuite des sous-échantillons plus larges ($>$ 5 000 points) pour les tests prédictifs *a posteriori*. Mais dans notre cas, dû à des contraintes de temps ($\approx$ 30 minutes par point) nous n'avons pas pu obtenir des sous-échantillons d'aussi bonne qualité ; des sous-échantillons de seulement 100 points ont été utilisés pour les analyses prédictives *a posteriori* (tableau 5).

Avec $MM6_{cov}$, les valeurs p sont visiblement plus congruentes entre deux chaînes MCMC mais restent néanmoins difficilement interprétables ; la moyenne des valeurs p

sur les $P_j(\Xi)$, avant la filtration des modulations covarion (figure 37b), est de beaucoup supérieure à la moyenne des valeurs p sur les $Z_j(\Xi)$ (figure 38b)(tableau 11). Le phénomène inverse est observé lorsque les modulations covarion sont filtrées (figures 37c et 38c). Nous n'avons pas de réponse formelle permettant d'expliquer ces résultats. Il s'agirait en fait d'un détail d'implémentation non identifié qui semble biaiser les distributions prédictives *a posteriori* des $P_j(\Xi)^s$ non filtrés et des $Z_j(\Xi)^s$ filtrés par rapport aux distributions observées correspondantes.

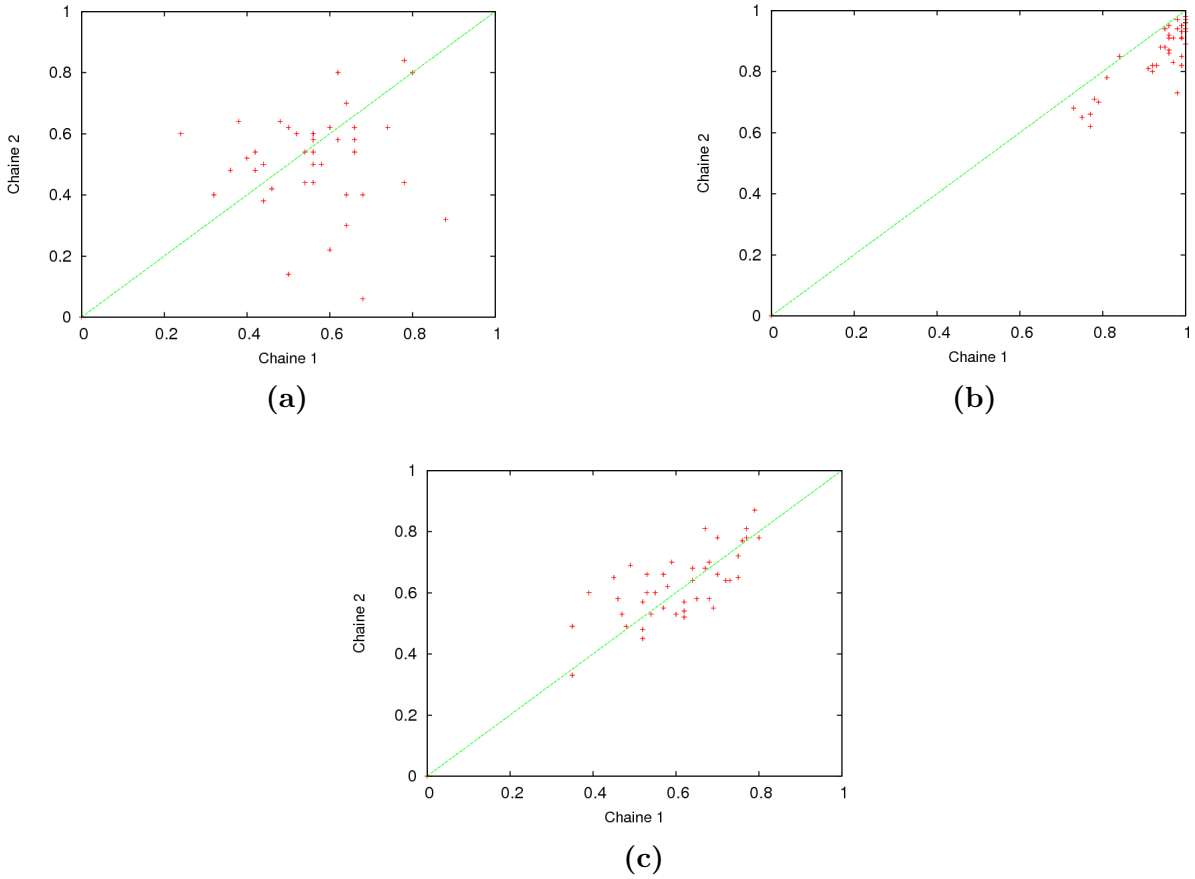


Figure 37 – Tests de congruence des valeurs p prédictives *a posteriori* sur les $P_j(\Xi)$. **a** Sous le modèle $MM6_{\Gamma}$. **b** Sous le modèle $MM6_{cov}$. **c** Sous le modèle $MM6_{cov}$ après filtration des modulations covarion.

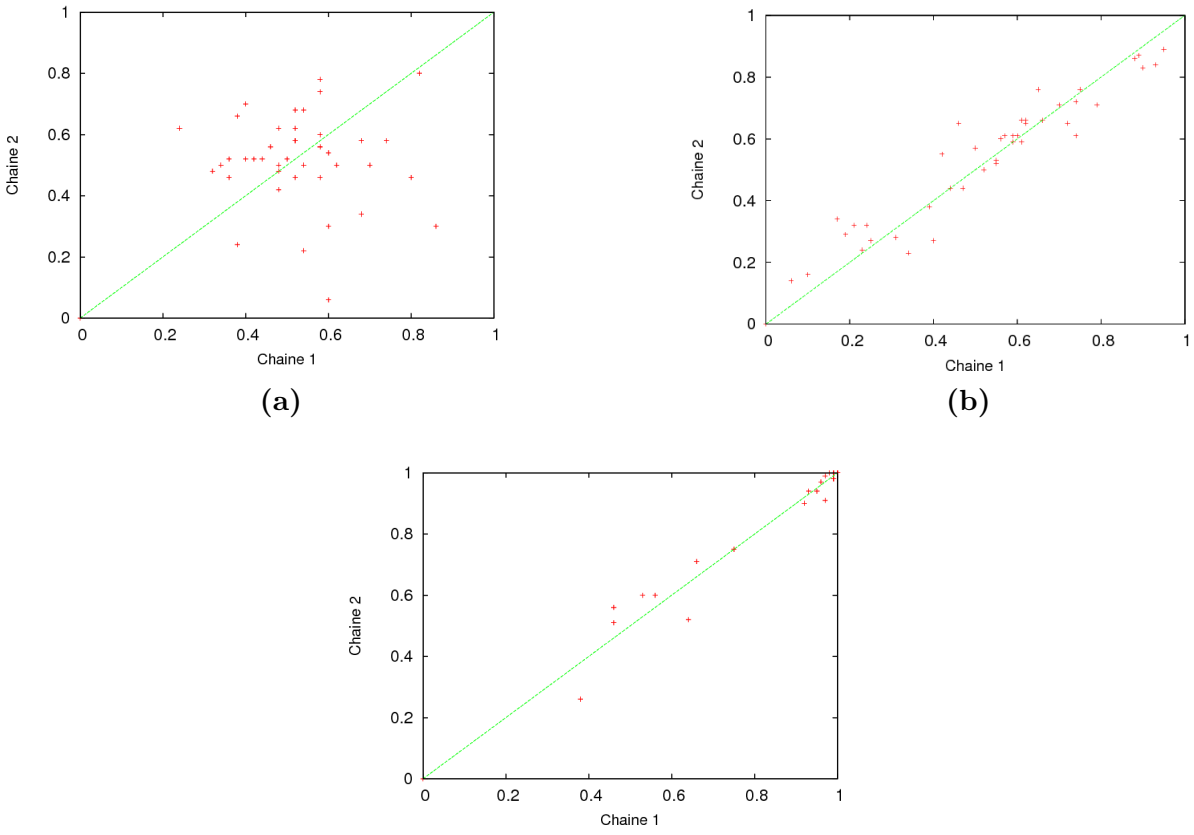


Figure 38 – Tests de congruence des valeurs p prédictives *a posteriori* sur les $Z_j(\Xi)$. **a** Sous le modèle $MM6_\Gamma$. **b** Sous le modèle $MM6_{cov}$. **c** Sous le modèle $MM6_{cov}$ après filtration des modulations covarion.

Tableau 11 – Moyennes et variances des valeurs p prédictives *a posteriori*

		Statistiques			
		Tests sur $P_j(\Xi)$		Tests sur $Z_j(\Xi)$	
Model	Chaînes	Moyenne	Variance	Moyenne	Variance
$MM6_\Gamma$	1	0.56	0.018	0.53	0.019
$MM6_\Gamma$	2	0.51	0.025	0.52	0.022
$MM6_{cov}$	1	0.94	0.0067	0.53	0.054
$MM6_{cov}$	2	0.86	0.0099	0.54	0.043
$MM6_{cov}{}^a$	1	0.61	0.014	0.91	0.033
$MM6_{cov}{}^a$	2	0.62	0.012	0.91	0.033

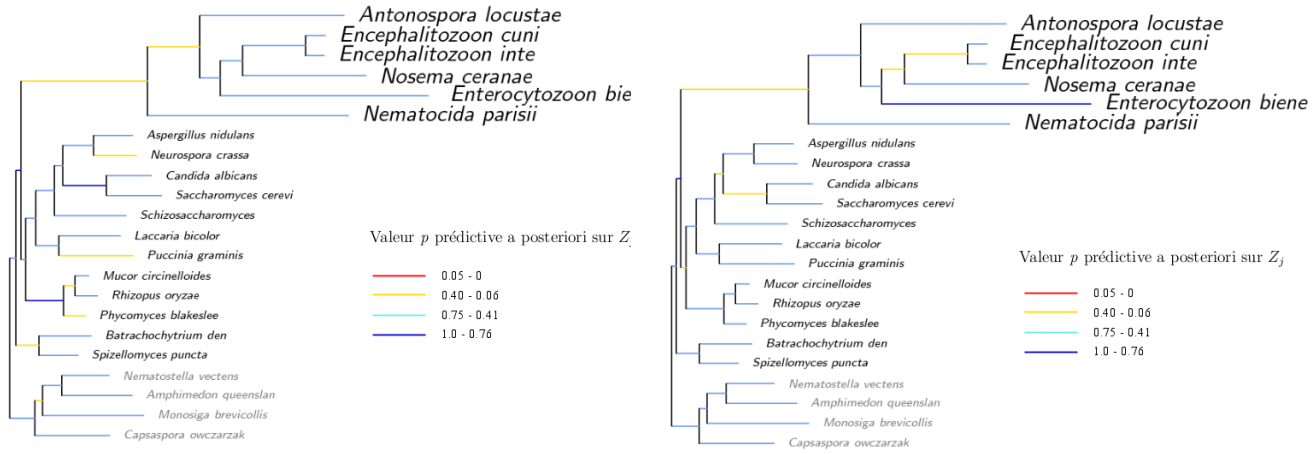
[a] Après filtration des modulations covarion

La suite logique aux tests de congruence des valeurs p c'est de visualiser celles-ci directement sur les branches d'arbre du jeu de données de microsporidies. Dans la suite, nous avons choisi de concentrer nos analyses sur la statistique $Z_j(\Xi)$ puisqu'elle est statistiquement plus significative que $P_j(\Xi)$.

Avec le modèle $MM6_\Gamma$, nous avons vu que la congruence des valeurs p est nettement insuffisante pour y détecter des branches enrichies en DF. Nous avons néanmoins présenté les arbres colorés correspondants sur la figure 39 ; la couleur respective de chacune des branches correspond à l'intervalle dans lequel sa valeur p sur $Z_j(\Xi)$ se situe selon la légende qui est jointe à la figure. Il est ainsi visuellement plus aisé de conclure que sous ce modèle rien d'intéressant ne ressort avec cette analyse prédictive *a posteriori*. Du moins avec les conditions présentes d'échantillonnage et de sous-échantillonnage (tel que nous avons discuté plus haut).

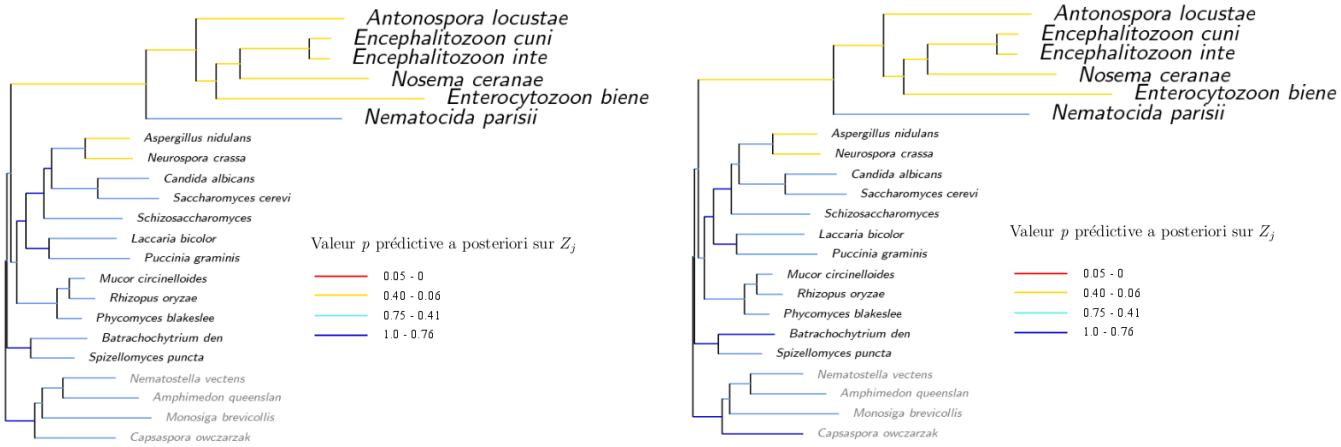
Il semble encore ici que le modèle $MM6_{cov}$ soit plus performant que $MM6_\Gamma$ pour détecter des branches enrichies en DF. En effet, malgré qu'aucune branche n'ait de valeur p sur les $Z_j(\Xi)$ qui est inférieure à 0.05, uniquement celles à l'intérieur du groupe des microsporidies ont des valeurs p inférieures ou égales à 0.75 lorsque les modulations covarion sont filtrées (figure 41). Les branches de ce groupe sont également mises en évidence avant la filtration des modulations covarion (avec des valeurs p inférieures ou égales à 0.40) mais de façon moins spécifique car deux autres branches à l'extérieur du groupe ressortent ; l'une à la base de *Aspergillus nidulans* et l'autre à la base de *Neurospora crassa*.

À la lumière de nos résultats d'analyses d'histoires substitutionnelles tirées de la distribution postérieure sous MM_{cov}, le phénomène d'hétérotachie contribuerait à mettre en évidence les branches sur lesquelles il y a une fréquence plus élevée de DF de type II (profil-spécifiques) (considérant également les résultats de $P_j(\Xi)$ filtrés présentés sur la figure 33c). Mais rappelons que la conception de MM_{cov} n'est pas optimale pour autant ; la non-convergence des chaînes MCMC, exagération du phénomène d'hétérotachie ainsi que le défaut de mémorisation de l'identité de l'état caché modulant dans l'état OFF. Cette configuration est donc sujette à quelques ajustements. D'autre part, il serait également possible d'implémenter un modèle avec des effets par branches (tel que déjà brièvement discuté plus haut). C'est à dire introduire des paramètres branche-spécifiques ajustant *a posteriori* la proportion transition entre états cachés/transitions entre états observés. Il serait ainsi possible de calculer et de comparer directement les distributions postérieures de ces paramètres sans passer par les histoires substitutionnelles.



(a) **(b)**

Figure 39 – Arbres colorés du jeu de données de microsporidies. Les couleurs de branches dépendent de leur valeur p prédictive *a posteriori* sur la statistique $Z_j(\Xi)$ sous le modèle $MM6_\Gamma$. **a** et **b** sont les résultats issus de deux chaînes MCMC indépendantes. Les noms d'espèces écrits en grands caractères correspondent au microsporidies, ceux écrits en caractères noirs de plus petit format sont les autres champignons et ceux en gris constituent le *outgroup*.



(a) **(b)**

Figure 40 – Arbres colorés du jeu de données de microsporidies. Les couleurs de branches dépendent de leur valeur p prédictive *a posteriori* sur la statistique $Z_j(\Xi)$ sous le modèle $MM6_{cov}$. **a** et **b** sont les résultats issus de deux chaînes MCMC indépendantes. Les noms d'espèces écrits en grands caractères correspondent au microsporidies, ceux écrits en caractères noirs de plus petit format sont les autres champignons et ceux en gris constituent le *outgroup*.

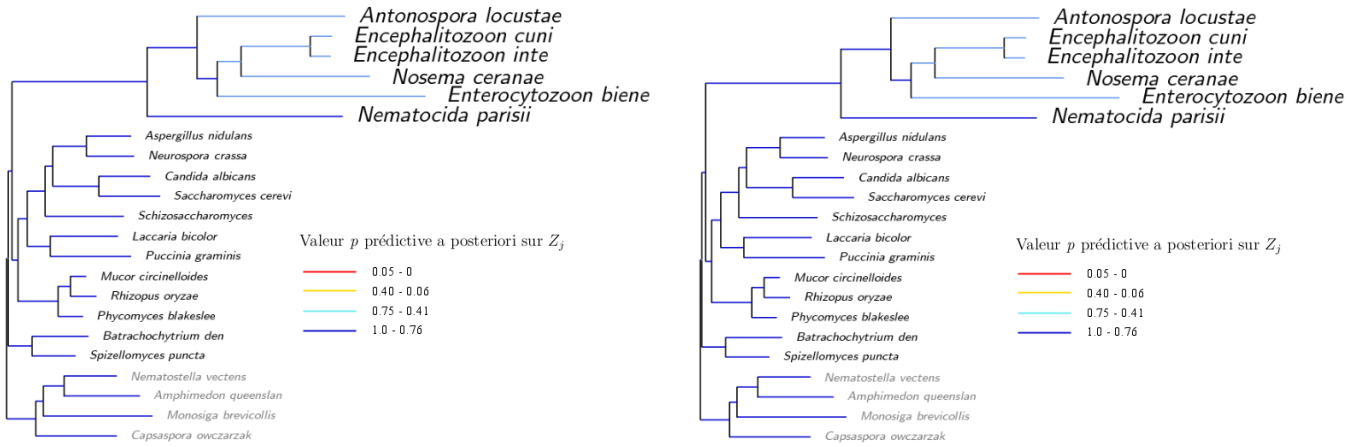


Figure 41 – Arbres colorés du jeu de données de microsporidies. Les couleurs de branches dépendent de leur valeur p prédictive *a posteriori* sur la statistique $Z_j(\Xi)$ sous le modèle $MM6_{cov}$ après la filtration des modulations covarion. **a** et **b** sont les résultats issus de deux chaînes MCMC indépendantes. Les noms d'espèces écrits en grands caractères correspondent au microsporidies, ceux écrits en caractères noirs de plus petit format sont les autres champignons et ceux en gris constituent le *outgroup*.

Pour notre deuxième test prédictif *a posteriori* visant aussi à évaluer la sensibilité de notre modèle aux DF, nous nous sommes inspiré des travaux de Roure & Philippe (2011) déjà introduits précédemment. Rappelons que leur approche consistait à démontrer qu'il est possible de mettre en évidence l'hétérogénéité temporelle des processus de substitution profil-spécifiques (hétéropécillie) directement dans les données empiriques. En résumé, l'idée est d'estimer la probabilité *a posteriori* PIP_n (*Probability of Identical Profil over n clades*) qu'un site soit affilié à différents profils CAT à travers un nombre n de clades. Un PIP_n près de 0 indiquant qu'il existe à ce site un fort signal d'hétéropécillie.

Nous avons tenté de reproduire un tel test mais adapté au processus Markov-modulé et dans un contexte d'analyse prédictive *a posteriori*. Pour ce faire, nous avons premièrement simulé des histoires substitutionnelles contraintes aux données observées et non contraintes tel que décrit précédemment. Ensuite, dans les deux cas un score $\wp_i$ est calculé pour chacun des sites i. Ce score correspond à la proportion d'espèces du groupe des microsporidies qui, au site i, ont leur acide aminé respectif dans un état caché différent de celui dans lequel se trouvent les acides aminés de toutes les autres espèces de l'arbre (l'état caché doit être le même à travers ces autres espèces sans quoi $\wp_i$ est automatiquement égal à 0).

La figure 42 présente un alignement hypothétique de séquences d'états cachés permettant de bien saisir la méthode de calcul de ces $\wp_i$. Ce score moyenné sur l'ensemble des sites permet d'obtenir $\mathcal{F}$ (équation 56), qui est la statistique utilisée pour cette analyse prédictive *a posteriori*. La valeur p est caculée en comptant le nombre de simulations parallèles pour lesquelles $\mathcal{F}^s \geq \mathcal{F}^o$. Celle-ci doit être inférieure à 0.05 pour rejetter l'hy-

pothèse de nullité voulant qu'il n'y ait aucun signal d'hétéropécillie dans le groupe des microsporidies.

Puisque nous nous intéressions ici au DF de type II, l'analyse a été faite exlusivement avec $\mathrm{MM3}_\Gamma$ et $\mathrm{MM6}_\Gamma$. Le tableau 12 présente les résultats de valeurs p obtenus sous chacun de ces deux modèles avec deux chaînes MCMC indépendantes. Comme cela a été le cas avec la statistique $Z_j(\Xi)$, les résultats sont non concluants ; non congruence sous $\mathrm{MM6}_\Gamma$ et valeurs p élevées sous $\mathrm{MM3}_\Gamma$. Les raisons sont probablement les mêmes que celles déjà invoquées ; la mauvaise qualité des chaînes MCMC et des conditions de sous-échantillonnage. La méthode de calcul des $\wp_i$ pourrait aussi être mise en cause. C'est à dire que la condition de l'état caché unique pour les espèces extérieures au groupe des microsporidies est peut-être trop rigide et entraîne une perte de signal.

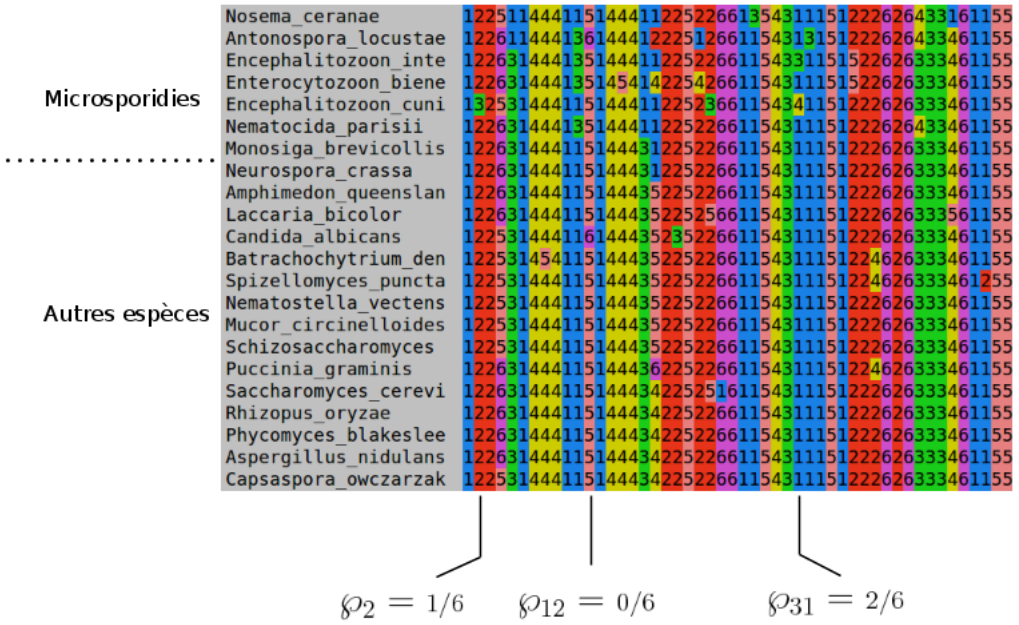


Figure 42 – Détection de DF aux feuilles de l'arbre à l'intérieur du groupe des microsporidies. Calcul des $\wp_i$ pour l'analyse prédictive *a posteriori* à partir d'un alignement hypothétique de séquences d'états cachés. Chacun des nombres correspond à l'un des 6 états cachés de *HSp6* et sont différenciés par un code de couleur.

Tableau 12 – Valeurs *p* prédictives *a posteriori* sur la statistique $\mathcal{F}$ visant à saisir des signaux d'hétéropécillie dans le groupe des microsporidies sous les modèles $MM3_\Gamma$ et $MM6_\Gamma$

Chaînes	Modèles	Valeurs p prédictives *a posteriori* sur $\mathcal{F}$
1	$MM6_\Gamma$	0.86
2	$MM6_\Gamma$	0.14
1	$MM3_\Gamma$	0.67
2	$MM3_\Gamma$	0.65

5 DISCUSSION

5.1 Évaluation d'ensemble des modèles MM_Γ et MM_{cov}

Nous avons conçu un modèle Marko-modulé pour accommoder les changements dans les préférences en acides aminés site-spécifiques au cours du temps. Il a été clairement démontré que de tels changements font partie du processus substitutionnel et que les modèles d'évolution moléculaire ont tout intérêt à les considérer (Holmes & Rubin 2002, Whelan 2008, Roure & Philippe 2011, Gaston et al. 2011). Des modèles Markov-modulés covarions pour les changements de vitesse d'évolution site-spécifique (hétérotachie) ont déjà été introduits en phylogénie moléculaire (Tuffley & Steel 1998, Galtier 2001, Huelsenbeck 2002, Wang et al. 2007). Ces modèles accommodent également d'une manière ou d'une autre le phénomène RAS et sont généralement mieux ajustés aux données que les modèles homogènes en temps. Notre modèle s'inspire jusqu'à un certain point des modèles covarions et des modèles Markov-modulés profil-spécifiques de Holmes & Rubin (2002) et de Whelan (2008). Mais fondamentalement, le notre a été développé sur une base d'alignements protéiques avec un mélange empirique de profils d'acides aminés. Vu sous cet angle, il peut être vu comme un modèle CAT paramétrique Markov-modulé. Autre nouvel apport de notre modèle par rapport aux autres approches Markov-modulées, est la non-reversibilité en temps qui offre potentiellement un meilleur ajustement aux données observées.

Les estimations *a posteriori* de $P(Q)$, $P(\Xi)$ et des paramètres cachés ($\eth$, η) sous les modèles $MM6_\Gamma$ et $MM3_\Gamma$ semblent montrer que notre approche Markov-modulée possède les bases nécessaires pour accommoder les phénomènes de divergence fonctionnelle ou plus généralement d'hétéropécillie. En moyenne, sous ces configurations, il y a à chacun des sites, 1 modulation et 20 transitions entre acides aminés à travers l'arbre avec le jeu de données de microsporidies (selon le calcul de $P(Q)$). Ce qui, qualitativement, est en accord avec la réalité biologique considérant que les phénomènes de divergences fonctionnelles et de changements de contextes environnementaux sont des évènements plus rares que les substitutions. De plus, ces proportions semblent effectivement extraites du signal contenu dans les données et non de la prior que nous avons choisie de façon à ce que le taux moyen de transition entre états cachés soit le même que celui entre états observés (équation 41).

Nous avons également noté que, autant sous MM_Γ que sous MM_{cov}, c'est l'état caché neutre (numéro 6), dit " poubelle ", qui a la fréquence d'équilibre la plus élvée. Cela semble suggérer que la richesse substitutionnelle de notre modèle est insuffisante. Autrement dit, lorsqu'un site ne trouve pas le profil qui lui convient, il choisit celui de l'état caché 6 par défaut. D'après les résultats de Lartillot & Philippe (2004) avec le modèle CAT non-paramétrique, au moins 25 profils seraient distribués à travers un jeu de données nucléaire eucaryote constitué d'aussi peu que 627 sites (versus 25 640 avec le jeu de données de microsporidies). Les résultats de Holmes & Rubin (2002) avec leur modèle Makov-modulé vont également dans le même sens : augmentation de la vraisemblance avec le nombre d'états cachés.

C'est principalement des complications d'ordre computationnelle qui nous ont contraint à restreindre à 6 le nombre d'états cachés. L'importante complexité algorithmique liée au calcul de la vraisemblance (algorithme de *pruning* + exponentielle de la matrice Q) fait que la performance de l'échantillonnage de la distribution postérieure par MCMC est rapidement réduite avec l'augmentation du nombre d'états cachés. D'autre part, le nombre d'états cachés a aussi un impact sur l'estimation *a posteriori* des taux relatifs d'échange entre états cachés. Alternativement, le calcul de l'espérance du nombre de transitions entre cachés par rapport au nombre total d'événements ($P(Q)$) permet d'atténuer cet impact sans toutefois nous convaincre qu'il n'y existe pas de problème de normalisation ailleurs. C'est pourquoi des efforts seront mis pour trouver précisement le détail d'implémentation qui en est responsable.

Ce qui ressort essentiellement des résultats obtenus avec la configuration covarion de notre modèle Markov-modulé (MM_{cov}) est que le nombre de transitions avec l'état caché OFF (numéro 0) est très élevé. Beaucoup plus élevé que le nombre de transitions entre les autres états cachés. Sous $MM6_{cov}$, il y en moyenne à chacun des sites 20 modulations pour 20 transitions entre acides aminés à travers l'arbre avec le jeu de données de microsporidies (selon le calcul de $P(Q)$). D'après ce que nous disent les histoires substitutionnelles (figure 35), la plupart de ces modulations impliquent l'état caché OFF. Il nous est difficile d'évaluer jusqu'à quel point le phénomène d'hétérotachie est exagéré ici. Mais toujours selon ces mêmes histoires substitutionnelles, parmi les modulations de type covarion, certaines semblent n'être que des objets mathématiques sans valeur (modulations parasites). Prenons par exemple la succession de modulations suivante ; $0 \succ 5 \succ 0 \succ 6 \succ 0$. Puisqu'il n'y a aucune transition entre acides aminés dans les états cachés 5 et 6, il s'agit en fait d'un scénario équivalent à une continuité dans l'état OFF.

Les taux élevés de transition avec l'état caché OFF sous MM_{cov} pourraient être un signal d'hétérotachie mais combiné avec un symptôme causé par l'absence de RAS. Whelan (2008) a observé un phénomène comparable avec son modèle appliqué sur un jeu de données nucléotidiques ; des taux de transition élevés entre l'état caché lent et l'état caché rapide (figure 9a). Lorsque RAS est réintroduit (figure 9b), le phénomène d'hétérotachie est nettement atténué. Il serait intéressant de vérifier comment se comporterait MM_{cov} si on le combinait avec la D-Gam. Où encore prendre exemple sur le modèle covarion de Huelsenbeck (2002) et de modéliser le phénomène d'hétérotachie de façon indépendante à celui de RAS. Avec cette approche, les vitesses de transitions entre les états cachés profil-spécifiques deviennent elles aussi indépendantes des taux entre sites. Un tel modèle serait cependant très demandant sur plan computationnel puisqu'il faudrait contruire une matrice de type MM_{cov} pour chacun des 4 taux discrétisés de la D-Gam.

Il est également intéressant de noter que sans RAS, le signal temporel profil-spécifique semble perturbé. En effet, les taux relatifs d'échange élevés observés avec RAS (figure 20a : $\eth_{15}$, $\eth_{26}$ et $\eth_{46}$) ne sont pas les mêmes que ceux sans RAS (figure 28c,d : $\eth_{12}$). Whelan (2008) observe aussi que sans RAS l'hétérogénéité temporelle qualitative n'est à peu près

pas observable comparée à lorsque RAS est incorporé au modèle (figure 9a versus 9b). Soulignons cependant qu'avec le modèle de Whelan (2008), l'impact sur l'hétérogénéité temporelle profil-spécifique n'est pas aussi claire puisque chacun des états cachés de celui-ci sont non seulement paramétrés avec un vecteur de fréquences d'équilibre mais aussi avec un paramètre pour la vitesse d'évolution (μ) et un troisième pour le taux de transition versus taux de transversion (κ). Néanmoins, que ce soit sous notre modèle Markov-modulé ou celui de Whelan (2008), tenter de reproduire l'effet RAS avec le phénomène de l'hétérotachie a définitivement un impact sur la capture des autres signaux d'hétérogénéité temporelle.

L'évaluation de la capacité de notre modèle à détecter les branches enrichies en divergences fonctionnelles (DF) dans le groupe des microsporidies va quelque peu à l'encontre de ce que nous venons de discuter à propos de l'impact de RAS sur la capture des signaux liés aux changements dans les préférences en acides aminés au cours du temps. En effet, nous avons pu noter que sous MM_{cov}, contrairement MM_{Γ}, la proportion de transitions entre états cachés par rapport au nombre total d'événements ($P_j(\Xi)$ calculés à partir des histoires substitutionnelles (Ξ) tirées de la distribution postérieure) semble plus élevée sur les branches du groupe des microsporidies. Les analyses prédictives *a posteriori* avec la statistique normalisée $Z_j(\Xi)$ abondent dans le même sens mais sous toute réserve étant donné l'important écart entre la distribution observée ($Z_j(\Xi)^o$) et la distribution non-contrainte ($Z_j(\Xi)^s$). Néanmoins, ce chapitre de notre projet mérite d'être éventuellement exploré plus en détail. Il serait premièrement essentiel de comprendre pourquoi un tel signal émerge du groupe des microsporidies plus particulièrement lorsque les modulations de type covarion sont filtrées.

Des analyses visuelles et statistiques d'histoires substitutionnelles détaillées (telles que celles apparaissant aux figures 35 et 36) à plus grande échelle pourraient nous aider à mieux comprendre ce qui se passe exactement sur les branches du groupes des microsporidies sous MM_{cov}. Il n'est pas évident à partir de nos résultats actuels d'identifier *a posteriori* quelles sont les successions d'événements de modulation qui sont proportionnellement distinctes sur ces branches par rapport aux autres branches de l'arbre ; est-ce des successions de type $k \succ l$ (modulations correspondant à des changements temporels de préférence en acides aminés) ou encore de type $k \succ 0 \succ l$ (correspondant à un signal d'hétérotachie combiné à un signal de changement de préférence en acides aminés) ? D'après la figure 36, le second type de successions semble être très fréquent sur la branche à la base du groupe des microsporidies. Mais étant donné que les taux de transition avec l'état caché OFF sont élevés ($C^{k,0}$ et $C^{0,l}$ sur la figure 27a) comparativement aux taux entre les états cachés profil-spécifiques, nous pouvons imaginer que cette succession est tout aussi fréquente à travers l'ensemble des branches de l'arbre. D'autre part, encore faudrait-il élaborer des méthodes de filtration plus raffinées. C'est à dire axer l'analyse davantage sur les successions de type k[transitions entre acides aminés] $\succ$ l[transitions entre acides aminés] en filtrant toutes celles à l'intérieur desquelles aucune transition entre acides aminés ne prend place dans les états cachés profil-spécifiques (succession de modulations parasites). Rappelons que les histoires substitutionnelles apparaissant sur les figures 34, 35, 36 représentent seulement ce qui se passe sur une seule branche. Elles ne nous informe pas sur l'histoire substitutionnelle

sur la branche précédente à l'intérieur de l'état caché à la base de la branche actuelle ni sur l'histoire substitutionnelle sur la branche suivante à l'intérieur de l'état caché à l'extrémité de la branche actuelle. Il faudrait donc idéalement afficher pour chacun des sites l'histoire substitutionnelle complète de l'arbre.

5.2 Approche Bayésienne et processus Markov-modulé

Pour des raisons que nous avons déjà mentionnées, la contrainte temps a été durant ce projet un obstacle majeur auquel nous avons dû faire face. La qualité des chaînes MCMC en a ainsi forcément subit les conséquences mais sans pour autant nous empêcher d'obtenir une vue d'ensemble du comportement des modèles MM_{Γ} et MM_{cov}. Pour obtenir des échantillons représentatifs de la distribution postérieure ciblée avec la méthode MCMC, il est essentiel de passer à travers un nombre de cycles suffisamment large (Norris 1998). Dans tous les cas, beaucoup plus large que ceux qui ont générés nos échantillons postérieurs à partir desquels nous avons procédé à diverses analyses Bayésiennes (soit en moyenne 3 000 cycles (tableau 5)). À titre d'exemple, pour étudier le comportement de leur modèle CAT non paramétrique, Lartillot & Philippe (2004) ont échantillonné 600 000 réalisations de θ de la distribution postérieure pour ensuite sous-échantillonner parmi les 500 000 points suivant la phase de *burn-in* un total de 10 000 points régulièrement espacé pour les estimations de paramètres *a posteriori*.

Les méthodes Bayésiennes, comparativement aux méthodes ML, offrent la possibilité d'élaborer des modèles d'évolution moléculaire plus complexes. De plus, elles procurent une mesure naturelle d'incertitude sur l'estimation des valeurs de paramètres inférées. Durant ce projet nous n'avons pas pleinement tiré avantage de ces possibilités. Afin d'étudier le comportement de notre modèle, il était essentiel de laisser libre ses paramètres cachés que sont les taux relatifs d'échange entre les états cachés ($\eth_{kl}$) ainsi que leurs fréquences d'équilibre (η_k). Nous avons d'autre part utilisé des données empiriques pour la topologie de l'arbre, les taux relatifs d'échange entre acides aminés et les profils des états cachés. Il aurait été intéressant d'estimer les profils ainsi que leur nombre directement à partir des données. Mais s'aventurer dans cette avenue dans les circonstances actuelles est à toutes fins pratiques inutile puisque nous ne parvenons pas préalablement à faire converger complètement les $\eth_{kl}$. Ce sont en fait eux qui ralentissent la convergence de la vraisemblance avec deux chaînes MCMC indépendantes. Sous $MM6_{cov}$, après 2100 points, les taux relatifs $\eth_{40}$ et $\eth_{50}$ ne sont d'ailleurs clairement pas congruents (figure 28a versus figure 28b). Cette difficulté est elle-même possiblement due à la faible richesse substitutionnelle à laquelle nous avons contraint notre modèle. Les mouvements Métropolis-Hasting faits à partir des données augmentées pourraient également contribuer à ralentir la congruence des $\eth_{kl}$ puisqu'il est probable qu'il y existe une forte dépendance entre ceux-ci et les histoires substitutionnelles. Autrement dit, $\eth$ et Ξ persistent à se retenir **mutuellement** hors de la distribution postérieure ciblée conditionnelle aux données observées. Une solution possible dans ce cas serait de combiner les mouvements en mode augmenté à des mouvements en mode *pruning*. Ces derniers ont une complexité algorithmique plus élevée

mais pourraient finalement être plus profitables. C'est une alternative réaliste tout à fait envisageable.

Une autre alternative possible pour optimiser notre échantillonneur Bayésien serait de définir des profils de plus petite dimension. Chaque profil aurait une propriété physico-chimique prédéfinie et serait ainsi conçu de manière à ce que le processus Bayésien puisse ajuster la fréquence d'équilibre d'uniquement les acides aminés les plus représentifs de cette propriété. Donc par exemple, si chaque profil était défini par seulement 4 acides aminés, une matrice Q avec 6 états cachés serait de degré 24 au lieu de 120. Ce qui contribuerait largement à accélérer le processus d'échantillonnage. Avec ces profils spécialisés, il serait aussi possible d'enrichir le processus substitutionnel avec plus d'états cachés et d'optimiser le contenu du mélange et sa dimension par *reversible jump* (Green 1995).

Finalement, il est fondamental que des analyses prédictives *a posteriori* soient conduites sur des sous-échantillons (idéalement plusieurs sous-échantillons) suffisamment larges comportant des réalisations décorrélées (Norris 1998, Gelman et al. 1996). Étant donné la faible qualité de nos chaînes MCMC, ces conditions n'ont malheureusement pas pu être entièrement respectées. Ce qui visiblement s'est réflété sur la congruence des valeurs p prédictive *a posteriori*. En fait, les seuls résultats extraits de ces analyses qui sont qualitativement interprétables sont ceux obtenus avec l'analyse prédictive *a posteriori* sur la statistique $Z_j(\Xi)$ sous MM_{cov} suite à la filtration des modulations covarions (figures 41a et 41b). Sous $MM6_\Gamma$, les analyses prédictives *a posteriori* sont définitivement non concluantes, et ce autant qualitativement que quantitativement ; congruence des $P_j(\Xi)$ calculées à partir des histoires substitutionnelles tirées de la distribution postérieure (figure 33a) mais pas des valeurs p avec les statistiques $P_j(\Xi)$ et $Z_j(\Xi)$ (figures 37a et 38a) ni des valeurs p avec la statistique $\mathcal{F}$ (tableau 12).

5.3 Perspectives d'amélioration du modèle

Nous avons vu que Galtier (2001) et Huelsenbeck (2002) optent pour deux approches différentes pour accommoder RAS avec leur modèle Markov-modulé respectif. Sous le modèle de Galtier (2001), seule une proportion de sites sont hétérotaches et modulent entre les quatres taux discrétisés d'une D-Gam. Les autres sites évoluent de façon homogène en temps sous le modèle RAS habituel. Nous pourrions imaginer un modèle de type MM_{cov} plus raffiné inspiré du modèle covarion de Galtier (2001). Sous ce modèle, une fraction de sites évoluerait de façon totalement homogène en temps (même profil et même taux d'évolution à travers l'ensemble de l'arbre). Car après tout, analogiquement à ce que stipule Galtier (2001) concernant l'homogénéité temporelle de la pression sélective à certains sites, ce ne sont pas nécessairement tous les sites qui changent leur préférence en acides aminés au cours du temps. L'autre fraction serait constituée de sites qui évoluent avec la possibilité de moduler entre deux niveaux d'états cachés ; un premier constitué d'états cachés profil-spécifiques et un deuxième pour les états cachés taux d'évolution-spécifiques. Un site pourrait ainsi moduler entre des taux d'évolution différents (selon les taux discrétisés de la D-Gam) tout en demeurant sous le même processus de substitution profil-spécifique. La

figure 43 représente une ligne de la matrice Q Markov-modulée correspondante d'ordre 24 (4 nucléotides $\times$ 3 états cachés $\times$ 2 taux d'évolution). Une adénine (A) aurait la possibilité d'être substituée par l'un ou l'autre des trois autres nucléotides (C, G ou T) du profil 1, de moduler vers un autre profil (avec le même taux d'évolution) ou encore de moduler vers un autre taux d'évolution tout en restant dans le profil 1.

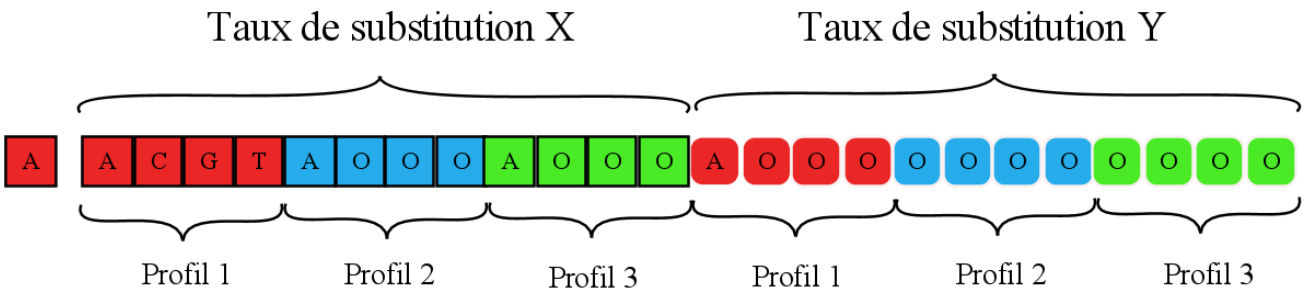


Figure 43 – Illustration d'un modèle Markov-modulé avec deux niveaux d'états cachés. Un premier niveau avec 3 profils et un second avec deux taux d'évolution. Ici, l'adénine (A) pourrait demeurer sous le même taux d'évolution, pour soit faire une transition dans le profil 1 ou moduler vers un autre profil, ou encore changer de taux d'évolution (dans le même profil). Un modèle nucléotidique est présenté pour simplifier.

Cette matrice à deux niveaux d'états cachés, contrairement à MM_Γ serait totalement indépendante de RAS. C'est à dire que seul les taux d'évolution à l'intérieur des profils (i.e. les vitesses de transitions entre acides aminés) changeraient. Les vitesses de modulation entre profils resteraient les mêmes peu importe l'état caché vitesse d'évolution-spécifique. Mais il serait également possible d'introduire des paramètres supplémentaires permettant d'ajuster les vitesses de modulations entre profils à l'intérieur de chacun des états cachés vitesse d'évolution-spécique. Ce qui permettrait d'estimer *a posteriori* ces paramètres et de comparer avec les résultats de Roure & Philippe (2011) suggérant que le phénomène d'hétéropécillie est positivement corrélé avec la vitesse de transition entre acides aminés.

Un désavantage considérable de MM_{cov} que nous avons déjà soulevé est que lorsque le processus module vers l'état caché OFF, il perd par la suite la mémoire de l'état caché profil-spécifique dans lequel il était auparavant. Ce qui est dépourvu de toute cohérence biologique. Pourquoi un site devrait-il forcément perdre toute préférence en acides aminés lorsque la pression sélective qui s'exerce sur lui est augmentée. Le processus Markov-modulé avec deux niveaux d'états cachés permettrait quant à lui d'être plus accommodant à ce niveau. Tel qu'illustré à la figure 43, si une position adénine dans le profil 1 module vers le taux de substitution Y, elle peut demeurer dans ce profil 1 et ensuite choisir oui ou non de changer son mode de substitution.

Nous en convenons qu'avec des profils consitués de 20 acides aminés chacun, une matrice Markov-modulée avec deux niveaux de modulation serait considérablement volumineuse. Mais au lieu de sommer quatre vraisemblances à chaque site comme c'est le cas sous MM_Γ

(un vecteur de vraisemblances conditionnelles pour chacun des quatres taux discrétisés de la D-Gam), il serait possible de diviser par deux cette charge computationnelle ; une vraisemblance conditionnelle à un processus Markov-modulé (pour la proportion de sites sous le processus substitutionnel hétérogène en temps) sommée à la vraisemblance conditionnelle à un processus homogène en temps (pour la proportion résiduelle de sites, en fixant à 0 tous les taux de modulation de la matrice Q). Donc, deux calculs de vraisemblance à chaque site au lieu de quatre. De plus, l'utilisation de profils plus compacts prédéfinis avec des propriétés physico-chimiques spécifiques comme nous l'avons proposé plus haut, pourrait contribuer à reduire la complexité algorithmique.

L'autre amélioration qui pourrait être apportée à notre modèle s'inspire des travaux de Galtier & Gouy (1998) sur la modélisation de l'hétérogénéité du contenu G+C au cours du temps. Nous avons vu dans l'introduction que leur modèle autorise chaque branche j de l'arbre à optimiser son contenu G+C avec un paramètre π^j_{GC}. Une approche comparable serait souhaitable pour modéliser l'hétérogénéité temporelle de la proportion de transitions entre états cachés profil-spécifiques par rapport au nombre de transtions entre états observés. Rappelons qu'*a priori* sous MM_Γ et MM_{cov} cette proportion est uniforme à travers l'ensemble de l'arbre. Avec une approche effet par branche, il ne serait plus nécessaire de calculer des $P_j(\Xi)$ à partir d'histoires substitutionnelles tirées de la distribution postérieure. L'identification de branches potentiellement atypiques se ferait ainsi en calculant directement les espérances *a posteriori* des valeurs de paramètres d'effet par branche.

5.4 Modèles Markov-modulés et estimations topologiques

À ce jour, peu d'études ont clairement démontré que les modèles Markov-modulés avaient un impact sur la reconstruction de liens évolutifs entre espèces. Parmi celles connues qui ont tenté d'accommoder le phénomène d'hétérotachie à partir de jeux de données nucléotidiques, mentionnons le cas d'une phylogénie d'angiospermes à partir d'un jeu de données chloroplastiques (Wang et al. 2007), celui celui d'une phylogénie d'opisthocontes à partir d'un jeu de données nucléaires (Ruiz-Trillo et al. 2004) et celui d'une phylogénie de dinoflagellés avec un jeu de données plastidiques (Schalchian-Tabrizi et al. 2006). Avant même de songer à utiliser notre modèle Markov-modulés pour accommoder l'hétérogénéité temporelle des préférence en acides aminés site-spécifique, il faudra évidemment régler les problèmes d'optimisation dont nous avons discutés. Et idéalement, il faudrait aussi au préalable envisager la conception de modèle Markov-modulés plus raffinés tel celui que nous avons imaginé avec deux niveaux de modulations.

L'échantillonnage de topologies de la distribution postérieure impose une charge computationnelle supplémentaire à la méthode MCMC. Même s'il existe une approche parallélisée pour les mouvements de type SPR (*Subtree pruning and regrafting* (Felsenstein 2004)) implémentée par Lartillot et al. (2013), l'impact sur le temps nécessaire pour atteindre la zone de convergence augmente significativement avec le nombre d'éspèces présentes

dans l'arbre. Avec un modèle Markov-modulé non-réversible en temps comme le notre, il faut également songer aux mouvements sur l'emplacement de la racine.

L'objectif principal de la phylogénie moléculaire demeure fondamentalement celui de recontruire des phylogénies non résolues. Notre modèle n'est définitivement pas prêt pour affronter cette problématique. Mais il semble néanmoins qualitativement disposer d'un certain potentiel à saisir des signaux d'hétérogénéité temporel dans les modes substitutionnels. Les tests de performance suggèrent qu'il contribue légèrement (relativement à la contribution très élevée de RAS) à améliorer l'ajustement aux données observées comparativement aux modèles équivalents homogènes en temps (CAT6+Γ et CAT3+Γ). Mais encore ici un scepticisme scientifique persiste. Est-ce que le léger gain en ajustement signifie que notre modèle est à toute fin pratique un modèle homogène en temps ou si c'est parce que les changements temporels dans les modes substitutionnels expliquant les données observées sont négligeables par rapport aux autres formes d'hétérogénéité. D'autres tests de performance (avec d'autres jeux de données et d'autres configurations de modèles Markov-modulés) seront nécessaires pour répondre à ces questions.

6 BIBLIOGRAPHIE

Adachi, J. & Hasegawa, M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial dna. *J. Comput. Biol.* **42** : 459–468.

Adkins, R.M., Honeycutt, R.L. & Disotell, T.R. 1996. Evolution of eutherian cytochrome c oxidase subunit II : heterogeneous rates of protein evolution and altered interaction with cytochrome c. *Mol. Biol. Evol.* **13** : 1393–1404.

Akaike, H. 1977. On entropy maximization principle. In :Krishnaiah, P.R., Applications of Statistics, North-Holland, Amsterdam .

Atkinson, K. 1989. An Introduction to Numerical Analysis (2nd ed.), New York : John Wiley and Sons .

Bateman, A., Birney, E., Durbin, R., Eddy, S., Howe, K. & Sonnhammer, E. 2000. The pfam protein families database. *Nucl. Acids Res.* **28** : 263–266.

Berger, J.O. 1985. Statistical decision theory and Bayesian analysis. Berlin : Springer-Verlag .

Bergsten, J. 2005. A review of long-branch attraction. *Cladistics* **21** : 163–193.

Blanquart, S. & Lartillot, N.. 2008. A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* **25** : 842–858.

Blanquart, S. & Lartillot, N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23** : 2058–2071.

Bogatyreva, N., Finkelstein, A. & Galzitskaya, O. 2006. Trend of amino acid composition of proteins of different taxa. *Bioinform. Comput. Biol.* **4** : 597–608.

Bollback, J.P. 2005. Posterior mapping and posterior predictive distributions, In : Statistical methods in Molecular Evolution, Eds. R. Nielsen, Springer, New York pp. 439–462.

Boussau, B. & Gouy, M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* **55(5)** : 756–768.

Brinkmann, H. & Philippe, H. 1999. Archaea sister group of bacteria ? indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* **16** : 817–825.

Bruno, W. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.* **13** : 1368–1374.

Bryant, D., Galtier, N. & Poursat, M. 2005. Likelihood calculation in molecular phylogenetics. In O. Gascuel, *Mathematics of Evolution and Phylogeny*. Oxford University Press, New York .

Caffrey, B., Williams, T., Jiang, X., Toft, C., Hokamp, K. & Fares, M. 2012. Proteome-wide analysis of functional divergence in bacteria : Exploring a host of ecological adaptations. *PLos ONE* **7(4)** : e35659.

Capella-Gutirrez, S., Marcet-Houben, M. & Gabaldon, T. 2012. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BCM Biology* **10 : 47** : doi :10.1186/1741–7007–10–47.

Cavalli-Sforza, L.L. & Edwards, A.W. 1967. Phylogenetic analysis. Models and estimation procedures. *Am. J Hum Genet.* **19** : 233–257.

Chakrabarti, S., Bryant, S. & Panchenko, A. 2007. Functional specificity lies within the properties and evolutionary changes of amino acids. *J.Mol.Biol.* **373** : 801–810.

Das, S., Paul, S., Bag, S. & Dutta, C. 2006. Analysis of nanoarchaeum equitans genome and proteome composition : indications for hyperthermophilic and parasitic adaptation. *BCM Genomics* **7** : 1–16.

Dayhoff, M., Eck, R. & Park, C. 1972. A model of evolutionary change in proteins. Pp. 88–89 In M. Dayhoff, ed., Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington, D.C .

Dayhoff, M., Schwartz, R. & Orcutt, B. 1978. A model of evolutionary change in proteins. Pp. 345–352 In M. Dayhoff, ed., Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington, D.C .

Dempster, A., Laird, N. & Rubin, D. 1977. Maximum likelihood fron incomplete data via the EM algorithm. *J.Roy. Statistical Soc. sect. B* **39** : 1–38.

Farris, J.S. 1970. Methods for computing wagner trees. *Syst. Zool.* **19** : 83–92.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27** : 401–410.

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

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associate Inc. ,Sunderland, Mass .

Felsenstein, J. & Churchill, G.A. 1996. A hidden markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13** : 93–104.

Ferguson, T. 1973. Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1(2)** : 209–230.

Fitch, W. 1971a. The nonidentity of invariable positions in the cytochromes c of different species. *Bioche. Genet.* **5** : 231–241.

Fitch, W.M. 1971b. Toward defining the course of evolution : Minimal change for a specific tree topology. *Syst. Zool.* **20** : 406–416.

Fitch, W.M. & Margoliash, E. 1967. Construction of phylogenetic trees. *Science* **155** : 279–284.

Fitch, W.M. & Markowitz, E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4** : 579–593.

Foster, P. 2004. Modeling compositional heterogeneity. *Syst.Biol.* **53(3)** : 485–495.

Foster, P. & Hickey, D. 1998. Compositional bias may affect both dna-based and protein-based phylogenetic reconstructions. *J.Mol.Evol.* **48** : 284–290.

Foster, P., Jermiin, L. & Hickey, D. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J.Mol.Evol.* **44** : 282–288.

Gadagkar, S. & Kumar, S. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Mol. Biol. Evol.* **22** : 2139–2141.

Galtier, G. & Gouy, M. 1998. Inferring pattern and process : Maximum-likelihood implementation of a nonhomogeneous model of dna sequence evolution for phylogenetic analysis. *Mol.Biol.Evol.* **15(7)** : 871–879.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.* **18** : 866–873.

Galtier, N., Gascuel, O. & Jean-Marie, A. 2005. Markov Model in Molecular Evolution, In : Statistical methods in Molecular Evolution, Eds. R. Nielsen, Springer, New York pp. 4–24.

Galtier, N. & Jean-Marie, A. 2004. Markov-modulated Markov chains and the covarion process of molecular evolution. *J. Comput. Biol.* **11(4)** : 727–733.

Gascuel, O. 2000. On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Mol. Biol. Evol.* **17** : 401–405.

Gaston, D., Susko, E. & Roger, A.J. 2011. A phylogenetic mixture model for the identification of functionally divergent protein residues. *Bioinformatics* **27** : 2655–22663.

Gaucher, E. & Miyamoto, M. 2005. A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. *Mol. Phylogenet. Evol.* **37** : 928–931.

Geisser, S. & Eddy, W. 1979. A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** : 153–160.

Gelfand, A. 1995. Model determination using sampling-based methods, In : Markov Chain Monte Carlo In Practice, Eds. W. Gilks, S. Richardson and D. Spiegelhalter, Chapman Hall, London, pp. 145–161.

Gelman, A. 1998. Simulating normalizing constants : from importance sampling to bridge sampling to path sampling. *Stat.Sci.* **13** : 163–185.

Gelman, A., Meng, X. & Stern, H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.* **6** : 733–807.

Geman, S. & Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* **6** : 721–741.

Gilks, W. & Roberts, G. 1996. Improving MCMC mixing. In : Markov Chain Monte Carlo in practice, eds W.R.Gilks, S. Richardson and D. J. Spiegelhalter. London :Chapman and Hall .

Goldman, N., J.L., T. & Jones, D. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149** : 445–458.

Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** : 711–732.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol.Biol.Evol.* **16(12)** : 1664–1674.

Gu, X., Fu, Y. & W.H., L. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol.Biol.Evol.* **12** : 546–557.

Haight, F. 1967. Handbook of the Poisson Distribution, New York, John Wiley and Sons .

Halpern, A. & Bruno, W. 1998. Evolutionary distances for protein-coding sequences : modeling site-specific residue frequencies. *Mol. Biol. Evol.* **15** : 910–917.

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

Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** : 97–109.

Henikoff, D. & al. 1997. Gene families : the taxonomy of protein paralogs and chimeras. *Science* **278** : 609–614.

Holder, M. & Lewis, P. 2003. Phylogeny estimation : traditional and bayesian approaches. *Nat.Re.Genet.* **4** : 275–284.

Holmes, I. & Rubin, G.M. 2002. An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.* **317** : 2655–2663.

Hordijk, W. & Gascuel, O. 2005. Improving the efficiency of spr moves in phylogenetic tree search methods based on maximum likelihood. *Bioinormatics* **21** : 4338–4347.

Huelsenbeck, J. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* **44** : 17–48.

Huelsenbeck, J. & Hillis, D. 1993. Success of phylogenetic methods in the 4-taxon case. *Syst. Biol.* **42** : 247–264.

Huelsenbeck, J., Hillis, D. & Jones, R. 1996. Parametric Boostrapping in Molecular Phylogenetics : Applications and Performance. Wiley-Liss, New-York pp. 19–45.

Huelsenbeck, J., Larget, B., Miller, R. & Ronquist, F. 2002. Applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* **51(5)** : 673–688.

Huelsenbeck, J.P. 2002. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* **19** : 698–707.

Huelsenbeck, J.P., Rannala, B. & Larget, B. 2000. A bayesian framework for the analysis of cospeciation. *Evolution* **54** : 352–364.

Inagaki, Y., Susko, E., N.M., F. & Roger, A. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in ef-1alpha phylogenies. *Mol. Biol. Evol.* **21** : 1340–1349.

Jaynes, E. 2003. Probability theory. The logic of science. Cambridge University Press, Cambridge, U.K .

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.* **31** : 203–222.

Jeffreys, H. 1961. Theory of probability. *Oxford University Press* .

Jermiin, L., Graur, D., Lowe, R. & Crozier, R. 1994. Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. *J.Biol.Evol.* **39** : 160–173.

Johnson, N. & Kotz, S. 1969. Discrete Distributions, boston : Houghton Miin .

Jones, D., Taylor, W. & Thornton, J. 1992. The rapid generation of mutation data matrices from protein sequences. *Cabios* **8** : 275–282.

Jukes, T.H. & Cantor, C. 1969. Evolution of protein molecules. *Mammalian Protein Metabolism* **3** : 21–132.

Keeling, P.J. & Fast, N.M. 2002. Microsporidia : biology and evolution of highly reduced intracellular parasites. *Annu. Rev. Microbiol.* **56** : 93–116.

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

Kolaczkowski, B. & Thornton, J. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431** : 980–984.

Koshi, J. & Goldstein, R. 1995. Context-dependent optimal substitution matrices. *Protein Eng.* **8** : 641–645.

Koshi, J. & Goldstein, R. 2001. Analysing site heterogeneity during protein evolution. *Pacific Symposium on Biocomputing* **6** : 191–202.

Koshi, J.M. & Goldstein, R.A. 1998. Models of natural mutations including site heterogeneity. *Proteins* **32** : 289–295.

Larget, B. & Simon, D.L. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16(6)** : 750–759.

Lartillot, N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. *J. Comput. Biol.* **13** : 1701–1722.

Lartillot, N., Brinkmann, H., & Philippe, H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7(Suppl 1)** : S4.

Lartillot, N. & Philippe, H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21** : 1095–1109.

Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. 2013. Phylobayes mpi : Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62(4)** : 611–615.

Le, S. & Gascuel, O. 2008. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25(7)** : 1307–1320.

Lemieux, C., Otis, C. & Turmel, M. 2007. A clade uniting the green algae mesostigma viride and chlorokybus atmophyticus represents the deepest branch of the streptophyta in chloroplast genome-based phylogenies. *BMC Biology* **5** : 2.

Lewis, P.O., Xie, W., Chen, M.H., Fan, Y. & Kuo, L. 2013. Posterior predictive Bayesian phylogenetic model selection. *Syst. Biol.* **doi : 10.1093/sysbio/syt068**.

Li, S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Ph.D. Dissertation, Ohio State Univ., Columbus .

Li, W. 1983. Evolution of duplicated genes. In Nei,M. and Koehn,R.K. (eds) Evolution of genes and proteins. Sinauer Associates, Sunderland, MA pp. 14–37.

Lockhart, P., Larkum, A., Steel, M., Waddell, P. & Penny, D. 1996. Evolution of chlorophyll and bacteriochlorophyll : the problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA* **93** : 1930–1934.

Lopez, P., Casane, D. & Philippe, H. 2002. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19** : 1–7.

Mau, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Ph.D. Dissertation, Univ. Wisconsin, Madison .

Meng, X. 1994. Posterior predictive p-values. *Ann. Stat.* **22** : 1142–1160.

Metropolis, N., Rosenbluth, W., A., Rosenbluth, M.N., Teller, A.H. & Teller, E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21** : 1087–1092.

Moler, C. & Van Loan, C. 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review* **45(1)** : 3–49.

Muto, A. & Osawa, S. 1987. The guanine and cytosine content of genomic dna and bacterial evolution. *Proc.Natl.Acad.Sci. USA* **84** : 166–169.

Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. *Ann. Rev.Genet.* **30** : 371–403.

Neyman, J. & Pearson, E. 1933. On the problem of the most efficient tests of statistical hypotheses. *Mathematical, Physical and Engineering Sciences* **231** : 694–706.

Nielsen, R. 2002. Mapping mutations on phylogenies. *Syst. Biol.* **51(5)** : 729–739.

Nilesen, R. 2005. Statistical Methods in Molecular Evolution, Rasmus Nielsen (Ed.) .

Norris, J. 1998. Markov chains, Cambridge University Press, New York .

Ogata, Y. 1989. A Monte Carlo method for high dimensional integration. *Numerishe Mathematik* **55** : 137–157.

Pace, C. & Scholtz, J. 1998. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75(1)** : 422–427.

Philippe, H., Germot, A. & Moreira, D. 2000. The new phylogeny of eukaryotes. *Curr. Opin. Genet. Dev.* **10** : 596–601.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* **5** : 50.

Quang, L., Gascuel, O. & Lartillot, N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24(20)** : 2317–2323.

Quang, L., Gascuel, O. & Lartillot, N. 2008b. Phylogenetic mixture models for proteins. *Phil.Trans.R.Soc.B.* **363** : doi :10.1098/rstb.2008.0180.

Quang, L.S., Gascuel, O. & Lartillot, N. 2008c. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24(20)** : 2317–2323.

Quang, L.S., Lartillot, N. & Gascuel, O. 2008d. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B* **363** : 3965–3976.

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.* **51** : 754–760.

Rannala, B. & Yang, Z. 1996. Probability distribution of molecular evolutionary trees : a new method of phylogenetic inference. *Mol. Biol. Evol.* **43** : 304–311.

Rodrigue, N., Philippe, H. & Lartillot, N. 2008. Uniformization for sampling realizations of markov processes : applications to Bayesian implementations of codon substitution models. *Bioinformatics* **24(1)** : 56–62.

Rodriguez-Ezpeleta, N., Philippe, H., Brinkmann, H., Becker, B. & Melkonian, M. 2007. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of mesostigma in the streptophyta. *Mol. Biol. Evol.* **24** : 723–731.

Roure, B. & Philippe, H. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evol. Biol.* **11** : 17.

Ruiz-Trillo, I., Inagaki, Y., Davis, L., Sperstad, S., Landfald, B. & Roger, A. 2004. Capsaspora owczarzaki is an independent opisthokont lineage. *Curr.Biol.* **14** : R946–R947.

Saitou, N. & Nei, M. 1987. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** : 406–425.

Schalchian-Tabrizi, K., Skanseng, M., Ronquist, F., Klaveness, D., Bachvaroff, T., Delwiche, C., Botnen, A., Tengs, T. & Jakobsen, K. 2006. Heterotachy processes in rhodophyte-derived second-hand plastid genes : implications for addressing the origin and evolution of dinoflagellate plastids. *Mol.Biol.Evol.* **23** : 1504–1515.

Schneider, R., deDaruvar, A. & Sander, C. 1997. The HSSP database of protein structure-sequence alignments. *Nucl. Acids Res.* **25(1)** : 226–230.

Schneider, T. & Stephens, R. 1990. Sequence logos : A new way to display consensus sequences. *Nucl. Acids Res.* **18** : 6097–6100.

Schwartz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6(2)** : 461–464.

Sinharay, S. & Stern, H. 2002. On the sensitivity of bayes factors to the prior distributions. *The American Statistician* **56(3)** : 196–201.

Smyth, P. 2000. Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comput.* **9** : 63–72.

Snir, M., Otto, S., Huss-Lederman, S., Walker, D. & Dongarra, J. 1995. MPI : The complete reference. MIT Press Cambridge, MA, USA .

Sokal, R. & Michener, C. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **4** : 1409–1438.

Spencer, M., Susko, E. & Roger, A. 2005. Likelihood, parsimony, and heterogeneous evolution. *Mol. Biol. Evol.* **22** : 1161–1164.

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

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

Thompson, J., Plewniak, F. & Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucl. Acids Res.* **27** : 2682–2690.

Thorne, J., Goldman, N. & Jones, D. 1996a. Combining protein evolution and secondary structure. *Mol.Evol.Biol.* **13** : 666–673.

Thorne, J., Goldman, N. & Jones, D. 1996b. Combining protein evolution and secondary structure. *Mol.Biol.Evol.* **13(5)** : 666–673.

Toft, C., Williams, T. & Fares, M. 2009. Genome-wide functional divergence after the symbiosis of proteobacteria with insects unraveled through a novel computational approach. *PLos ONE* **5(4)** : e1000344.

Tuffley, C. & Steel, M. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci.* **147** : 63–91.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Information Theory* **13** : 260–269.

Wang, H., Li, K., Susko, E. & Roger, A. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BCM Evol. Biol.* **8 :331**.

Wang, H.C., Spencer, M., Susko, E. & Roger, A.J. 2007. Testing for covarion-like evolution in protein sequences. *Mol. Biol. Evol.* **24** : 294–305.

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol. Biol. Evol.* **25(8)** : 1683–1694.

Whelan, S. & Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18** : 691–699.

Wilkinson, J. 1965. The Algebraic Eigenvalue Problem, Oxford University Press, Oxford .

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10** : 1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites : approximate methods. *J. Mol. Evol.* **39** : 306–314.

Yang, Z. & Rannala, B. 1997. Bayesian phylogenetic inference using dna sequences : a Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14** : 717–724.

Zhou, Y., Brinkmann, H., Rodrigue, N., Lartillot, N. & Philippe, H. 2010. A dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. *Mol. Biol. Evol.* **27(2)** : 371–384.